online issn: 1307-6892

irc 2022
XVI. international research conference
proceedings

open science index 16 2022

july 12-13, 2022 stockholm sweden

international scholarly and scientific research & innovation

Open Science

Open Science Philosophy

Open science encompasses unrestricted access to scientific research articles, access to data from public research, and collaborative research enabled by information and communication technology tools, models, and incentives. Broadening access to scientific research publications and data is at the heart of open science. The objective of open science is to make research outputs and its potential benefits available to the entire world and in the hands of as many as possible:

- Open science promotes a more accurate verification of scientific research results. Scientific inquiry and discovery can be sped up by combining the tools of science and information technologies. Open science will benefit society and researchers by providing faster, easier, and more efficient availability of research outputs.
- Open science reduces duplication in collecting, creating, transferring, and re-using scientific material.
- Open science increases productivity in an era of tight budgets.
- Open science results in great innovation potential and increased consumer choice from public research.
- Open science promotes public trust in science. Greater citizen engagement leads to active participation in scientific experiments and data collection.

Open Science Index

The Open Science Index (OSI) currently provides access to over thirty thousand full-text journal articles and is working with member and non-member organizations to review policies to promote and assess open science. As part of the open science philosophy, and by making open science a reality; OSI is conducting an assessment of the impact of open science principles and restructuring the guidelines for access to scientific research. As digitalization continues to accelerate science, Open science and big data hold enormous promise and present new challenges for policymakers, scientific institutions, and individual researchers.

OSI is helping the global scientific research community discover, evaluate, and access high-quality research output. Renowned for its editorially curated and refereed collection of the highest-quality publications, OSI has always been and will remain free-of-charge.

OSI provides an efficient and thorough discovery process to the open science research database and provides links and free access to full-text articles. There are 50 open access journal categories that are curated and refereed by international scientific committees, the in-house editorial team, and trusted partners. Since its inception in 2007, OSI has made more than thirty thousand peer-reviewed open access full-text journal articles (PDF versions) freely available online without cost, barriers, or restrictions.

Open Science Access

With the Open Science Index, researchers can discover and access trusted peer-reviewed open access full-text scientific research articles with confidence. OSI helps researchers find appropriate non-profit open access journals to publish their work.

OSI gives one-click access to online full-text PDFs and expands the reach to global society by giving users free access from anywhere around the globe. Through cutting-edge open science collaboration, in an innovative public partnership, the non-profit OSI is devoted to making science open and reusable.

To learn more, visit online at waset.org

Open Science

Open Society

An open society allows individuals to change their roles and to benefit from corresponding changes in status. Open science depends to a greater or lesser extent on digital technologies and innovations in structural processes by an open society. When realized, open science research and innovation can create investment opportunities for new and better products and services and therefore increase competitiveness and employment. Open science research and innovation is a key component of thematic open science priorities. Central to the open science digital infrastructure is enabling industry to benefit from digital technology and to underpin scientific advances through the development of an open society. Open science research and innovation can also contribute to society as a global actor because scientific relations can flourish even where global relations are strained. Open science has a critical role across many areas of decision making in providing evidence that helps understand the risks and benefits of different open science choices. Digital technology is making the conduct of open science and innovation more collaborative, more global, and more open to global citizens. Open society must embrace these changes and reinforce its position as the leading power for science, for new ideas, and for investing sustainably in the future.

It is apparent in open society that the way science works is fundamentally changing, and an equally significant transformation is taking place in how organizations and societies innovate. The advent of digital technology is making research and innovation more open, collaborative, and global. These exchanges are leading open society to develop open science and to set goals for research and innovation priority. Open science goals are materializing in the development of scientific research and innovation platforms and greater acceptance of scientific data generated by open science research. Open science research and innovation do not need help from open society to come up with great ideas, but the level of success ideas ultimately reach is undoubtedly influenced by regulation, financing, public support, and market access. Open society is playing a crucial role in improving all these success factors.

Open Science

Open science represents a new approach to the scientific process based on cooperative work and new ways of diffusing knowledge by using digital technologies and collaborative tools. These innovations capture a systemic change to the way science and research have been carried out for the last fifty years. Science is shifting from the standard practice of publishing research results in scientific publications after the research and reviews are completed. The shift is towards sharing and using all available knowledge at an earlier stage in the research process. Open science is to science what digital technology is to social and economic transactions: allowing end users to be producers of ideas, relations, and services and in doing so, enabling new working models, new social relationships and leading to a new modus operandi for science. Open science is as important and disruptive as e-commerce has been for the retail industry. Just like e-commerce, the open science research paradigm shift affects the whole business cycle of doing science and research. From the selection of research subjects to the carrying out of research, to its use and re-use, to the role of universities, and that of publishers are all dramatically changed. Just as the internet and globalization have profoundly changed the way we do business, interact socially, consume culture, and buy goods, these changes are now profoundly impacting how one does research and science.

The discussion on broadening the footprint of science and on novel ways to produce and spread knowledge gradually evolved from two global trends: Open Access and Open Source. The former refers to online, peer-reviewed scholarly outputs, which are free to read, with limited or no copyright and licensing restrictions, while open source refers to software created without any proprietary restriction and which can be accessed and freely used. Although open access became primarily associated with a particular publishing or scientific dissemination practice, open access already sought to induce a broader practice that includes the general re-use of all kinds of research products, not just publications or data. It is only more recently that open science has coalesced into the concept of a transformed scientific practice, shifting the focus of researchers' activity from publishing as fast as possible to sharing knowledge as early as possible. Open science is defined as the idea that scientific knowledge of all kinds should be openly shared as early as is practical in the discovery process. As a result, the way science is done in the future will look significantly different from the way it is done now. Open science is the ongoing evolution in the modus operandi of doing research and organizing science. This evolution is enabled by digital technology and is driven by both the globalization of the scientific community and increasing public demand to address the societal challenges of our times. Open science entails the ongoing transitions in the way research is performed, researchers collaborate, knowledge is shared, and science is organized.

Open science impacts the entire research cycle, from the inception of research to its publication, and on how this cycle is organized. The outer circle reflects the new interconnected nature of open science, while the inner circle shows the entire scientific process, from the conceptualization of research ideas to publishing. Each step in the scientific process is linked to ongoing changes brought about by open science, including the emergence of alternative systems to establish a scientific reputation; changes in the way quality and impact of research are evaluated; the growing use of scientific blogs; open annotation; and open access to data and publications. All institutions involved in science are affected, including research organizations, research councils, and funding bodies. The trends are irreversible, and they have already grown well beyond individual projects. Theses changes predominantly result from a bottom-up process driven by a growing number of researchers who increasingly employ social media in their research and initiate globally coordinated research projects while sharing results at an early stage in the research process.

Open science is encompassed in five schools of thought:

- the infrastructure school, concerned with technological architecture
- the public school, concerned with the accessibility of knowledge creation
- the measurement school, concerned with alternative impact assessment
- the democratic school, concerned with access to knowledge
- the pragmatic school, concerned with collaborative research

According to the measurement school, the reputation and evaluation of individual researchers are still mainly based on citation-based metrics. The h-index is an author-level metric that attempts to measure both the productivity and citation impact of the publications of a scientist or scholar. The impact factor is a measure reflecting the average number of citations to articles published in an academic journal and is used as a proxy for the relative importance of a journal.

Numerous criticisms have been made of citation-based metrics, primarily when used, and often misused, to assess the performance of individual researchers. These metrics:

- are often not applicable at the individual level
- do not take into account the broader social and economic function of scientific research
- are not adapted to the increased scale of research
- cannot recognize new types of work that researchers are performing

Web-based metrics for measuring research output, popularized as altmetrics, have recently received much attention: some measure the impact at the article level, others make it possible to assess the many outcomes of research in addition to the number of scientific articles and references. The current reputation and evaluation system has to adapt to the new dynamics of open science and acknowledge and incentivize engagement in open science. Researchers engaging in open science have growing expectations that their work, including intermediate products such as research data, will be better rewarded or taken into account in their career development. Vice-versa, the use, and reuse of open data will require appropriate codes of conduct requiring, for example, the proper acknowledgment of the original creator of the data.

These ongoing changes are progressively transforming scientific practices with innovative tools to facilitate communication, collaboration, and data analysis. Researchers that increasingly work together to create knowledge can employ online tools and create a shared space where creative conversation and collaboration can occur. As a result, the problem-solving process can be faster, and the range of problems that can be solved can be expanded. The ecosystem underpinning open science is evolving very rapidly. Social network platforms for researchers already attract millions of users and are being used to begin and validate more research projects.

Furthermore, the trends towards open access are redefining the framework conditions for science and thus have an impact on how open innovation is produced by encouraging a more dynamic circulation of knowledge. It can enable more science-based startups to emerge thanks to the exploitation of openly accessible research results. Open science, however, does not mean free science. It is essential to ensure that intellectual property is protected before making knowledge publicly available in order to subsequently attract investments that can help translate research results into innovation. If this is taken into account, fuller and broader access to scientific publications and research data can help to accelerate innovation. Investments that boost research and innovation in open science would benefit society with fewer barriers to knowledge transfer, open access to scientific research, and greater mobility of researchers. In this context, open access can help overcome the barriers that innovative organizations face in accessing the results of research funded by the public.

Open innovation

An open society is the largest producer of knowledge, but the phenomenon of open science is changing every aspect of the scientific method by becoming more open, inclusive, and interdisciplinary. Ensuring open society is at the forefront of open science means promoting open access to scientific data and publications alongside the highest standards of research integrity. There are few forces in this globe as engaging and unifying as science. The universal language of science maintains open channels of communication globally. Open society can maximize its gains through maintaining its presence at the highest level of scientific endeavor, and by promoting a competitive edge in the knowledge society of the information age. The ideas and initiatives described in this publication can stimulate anyone interested in open science research and innovation. It is designed to encourage debate and lead to new ideas on what and open society should do, should not do, or do differently.

An open society can lead to a research powerhouse; however, open society rarely succeeds in turning research into innovation and in getting research results to the global market. Open society must improve at making the most of its innovation talent, and that is where open innovation comes into play. The basic premise of open innovation is to open up the innovation process to all active players so that knowledge can circulate more freely and be transformed into products and services that create new markets while fostering a stronger culture of entrepreneurship. Open innovation is defined as the use of purposive inflows and outflows of knowledge to accelerate internal innovation. This original notion of open innovation was primarily based on transferring knowledge, expertise, and even resources from one company or research institution to another. This notion assumes that firms can and should use external ideas as well as internal ideas, and internal and external paths to market, as they seek to improve their performance. The concept of open innovation is continually evolving and is moving from linear, bilateral transactions and collaborations towards dynamic, networked, multi-collaborative innovation ecosystems. This means that a specific innovation can no longer be seen as the result of predefined and isolated innovation activities but rather as the outcome of a complex co-creation process involving knowledge flows across the entire economic and social environment. This co-creation takes place in different parts of the innovation ecosystem and requires knowledge exchange and absorptive capacities from all the actors involved, whether businesses, academia, financial institutions, public authorities, or citizens.

Open innovation is a broad term, which encompasses several different nuances and approaches. Two main elements underpin the most recent conceptions of open innovation: the users are in the spotlight and invention becomes an innovation only if users become a part of the value creation process. Notions such as user innovation emphasize the role of citizens and users in the innovation processes as distributed' sources of knowledge. This kind of public engagement is one of the aims of open science research and innovation. The term 'open' in these contexts has also been used as a synonym for 'user-centric'; creating a well-functioning ecosystem that allows co-creation and becomes essential for open innovation. In this ecosystem, relevant stakeholders are collaborating along and across industry and sector-specific value chains to co-create solutions for socio-economic and business challenges. One important element to keep in mind when discussing open innovation is that it cannot be defined in absolutely precise terms. It may be better to think of it as a point on a continuum where there is a range of context-dependent innovation activities at different stages, from research to development through to commercialization, and where some activities are more open than others. Open innovation is gaining momentum thanks to new large-scale trends such as digitalization and the mass participation and collaboration in innovation that it enables. The speed and scale of digitalization are accelerating and transforming the way one designs, develops, and manufactures products, the way one delivers services, and the products and services themselves. It is enabling innovative processes and new ways of doing business, introducing new cross-sector value chains and infrastructures.

Open society must ensure that it capitalizes on the benefits that these developments promise for citizens in terms of tackling societal challenges and boosting business and industry. Drawing on these trends, and with the aim of helping build an open innovation ecosystem in open society, the open society's concept of open innovation is characterized by:

- combining the power of ideas and knowledge from different actors to co-create new products and find solutions to societal needs
- creating shared economic and social value, including a citizen and user-centric approach
- capitalizing on the implications of trends such as digitalization, mass participation, and collaboration

In order to encourage the transition from linear knowledge transfer towards more dynamic knowledge circulation, experts agree that it is essential to create and support an open innovation ecosystem that facilitates the translation of knowledge into socio-economic value. In addition to the formal supply-side elements such as research skills, excellent science, funding and intellectual property management, there is also a need to concentrate on the demand side aspects of knowledge circulation, making sure that scientific work corresponds to the needs of the users and that knowledge is findable, accessible, interpretable and reusable. Open access to research results aims to make science more reliable, efficient, and responsive and is the springboard for increased innovation opportunities, e.g. by enabling more science-based startups to emerge. Prioritizing open science does not, however, automatically ensure that research results and scientific knowledge are commercialized or transformed into socio-economic value. In order for this to happen, open innovation must help to connect and exploit the results of open science and facilitate the faster translation of discoveries into societal use and economic value.

Open Science

Collaborations with global partners represent important sources of knowledge circulation. The globalization of research and innovation is not a new phenomenon, but it has intensified in the last decade, particularly in terms of collaborative research, international technology production, and worldwide mobility of researchers and innovative entrepreneurs. Global collaboration plays a significant role both in improving the competitiveness of open innovation ecosystems and in fostering new knowledge production worldwide. It ensures access to a broader set of competencies, resources, and skills wherever they are located, and it yields positive impacts in terms of scientific quality and research results. Collaboration enables global standard-setting, allows global challenges to be tackled more effectively, and facilitates participation in global value chains and new and emerging markets.

To learn more, visit online at waset.org

Open Science

Scholarly Research Review

The scholarly research review is a multidimensional evaluation procedure in which standard peer review models can be adapted in line with the ethos of scientific research, including accessible identities between reviewer and author, publishing review reports and enabling greater participation in the peer review process. Scholarly research review methods are employed to maintain standards of quality, improve performance, provide credibility, and determine suitability for publication. *Responsible Peer Review Procedure:* Responsible peer review ensures that scholarly research meets accepted disciplinary standards and ensures the dissemination of only relevant findings, free from bias, unwarranted claims, and unacceptable interpretations. Principles of responsible peer review:

- Honesty in all aspects of research
- Accountability in the conduct of research
- Professional courtesy and fairness in working with others
- Good stewardship of research on behalf of others

The responsibilities of peer review apply to scholarly researchers at all stages of peer review: Fairness, Transparency, Independence, Appropriateness and Balance, Participation, Confidentiality, Impartiality, Timeliness, Quality and Excellence, Professionalism, and Duty to Report.

Scholarly Research Review Traits:

- Scholarly Research Review Identities: Authors and reviewers are aware of each other's identity

- Scholarly Research Review Reports: Review reports are published alongside the relevant article

- Scholarly Research Review Participation: The wider academic community is able to contribute to the review process

- Scholarly Research Review Interaction: Direct reciprocal discussion between author(s) and reviewers, and/or between reviewers, is allowed and encouraged

- Scholarly Research Pre-review Manuscripts: Manuscripts are made immediately available in advance of any formal peer review procedures

- Scholarly Research Review Final-version Reviewing: Editorial revision of the language and format is conducted on the final version of the manuscript for publication

- Scholarly Research Review Platforms: The scholarly research review process is independent of the final publication of the manuscript and it is facilitated by a different organizational entity than the venue of publication

All submitted manuscripts are subject to the scholarly research review process, in which there are three stages of evaluation for consideration: pre-review manuscripts, chair-review presentation, and final-review manuscripts. All submitted full text papers, that may still be withstand the editorial review process, are presented in the conference proceedings. Manuscripts are tracked and all actions are logged by internal and external reviewers according to publication policy. External reviewers' editorial analysis consists of the evaluation reports of the conference session chairs and participants in addition to online internal and external reviewers' reports. Based on completion of the scholarly research review process, those manuscripts meeting the publication standards are published 10 days after the event date.

To learn more, visit online at waset.org

TABLE OF CONTENTS

Analyzing Sociocultural Factors Shaping Architects' Construction Material Choices: The Case of Jordan

Maiss Razem

***Abstract*—** The construction sector is considered a major consumer of materials that undergo processes of extraction, processing, transportation, and maintaining when used in buildings. Several metrics have been devised to capture the environmental impact of the materials consumed during construction using lifecycle thinking. Rarely has the materiality of this sector been explored qualitatively and systemically. This paper aims to explore socio-cultural forces that drive the use of certain materials in the Jordanian construction industry, using practice theory as a heuristic method of analysis, more specifically Shove et al. three-element model. By conducting semi-structured interviews with architects, the results unravel contextually embedded routines when determining qualities of three materialities highlighted herein; stone, glass and spatial openness. The study highlights the inadequacy of only using efficiency as a quantitative metric of sustainable materials and argues for the need to link material consumption with socio-economic, cultural, and aesthetic driving forces. The operationalization of practice theory by tracing materials' lifetimes as they integrate with competences and meanings captures dynamic engagements through the analyzed routines of actors in the construction practice. This study can offer policy makers better nuanced representation to green this sector beyond efficiency rhetoric and quantitative metrics.

***Keywords*—** architects' practices, construction materials, Jordan, practice theory

Maiss Razem is PhD Student with the University of Cambridge, Churchill College, CB3 0DS UK (phone: 962-79-9957942; e-mail: mjr216@ cam.ac.uk).

DEM Simulation of Crushable Pumice Sand

S. H. Bahmani, R. P. Orense

***Abstract*—**From an engineering point of view, pumice particles are problematic because of their crushability and compressibility due to their vesicular nature. Currently, information on the geotechnical characteristics of pumice sands is limited. While extensive empirical and laboratory tests can be implemented to characterize their behavior, these are generally time-consuming and expensive. These drawbacks have motivated attempts to study the effects of particle breakage of pumice sand through the Discrete Element Method (DEM). This method provides insights into the behavior of crushable granular material at both the micro- and macro-level. In this paper, the results of single particle crushing tests conducted in the laboratory are simulated using DEM through the open-source code YADE. This is done to better understand the parameters necessary to represent the pumice microstructure that governs its crushing features, and to examine how the resulting microstructure evolution affects a particle's properties. The DEM particle model is then used to simulate the behavior of pumice sand during consolidated drained triaxial tests. The results indicate the importance of incorporating particle porosity and unique surface textures in the material characterization and show that interlocking between the crushed particles significantly influences the drained behavior of the pumice specimen**.**

***Keywords*—**Pumice sand, triaxial compression, simulation, particle breakage.

Sayed Hessam Bahmani is with the University of Aucklanda, New Zealand (e-mail: sbah010@aucklanduni.ac.nz).

Gold–M Heterobimetallic Complexes: Synthesis and Initial Reactivity Studies

Caroline A. Rouget-Virbel, F. Dean Toste

Abstract**—** Heterobimetallic systems have been precedented in a wide array of bioinorganic and heterogeneous catalytic settings, in which cooperative bond-breaking and bond-forming events mediated by neighboring metal sites have been proposed but are challenging to study and characterize. Heterodinuclear transition-metal catalysis has recently emerged as a promising strategy to tackle challenging chemical transformations, including C−C and C−X couplings as well as small molecule activation. It has been shown that these reactions can traverse nontraditional mechanisms, reactivities, and selectivities when homo- and heterobimetallic systems are employed. Moreover, stoichiometric studies of transmetallation from gold complexes have demonstrated that R transfer from PPh_3–Au(I)R to Cp- and Cp*-ligated group 8/9 complexes is a viable elementary step. With these considerations in mind, we hypothesized that heterobimetallic Au–M complexes could serve as a viable and tunable catalyst platform to explore mechanism and reactivity. In this work, heterobimetallic complexes containing Au(I) centers tethered to Ir(III) and Rh(III) piano stool moieties were synthesized and characterized. Preliminary application of these complexes to a catalytic allylic arylation reaction demonstrates bimetallic cooperativity relative to their monomeric metal components.

Keywords**—**Bimetallic catalysis, gold(I), rhodium catalysis

Caroline A. Rouget-Virbel is with the Department of Chemistry, UC Berkeley, Berkeley, CA 94720 USA (phone: 609-356-8641; e-mail: caroline_rouget@berkeley.edu).

F. Dean Toste is with the Department of Chemistry, UC Berkeley, Berkeley, CA 94720 USA (e-mail: fdtoste@berkeley.edu).

Simulation of complex-shaped particle breakage using the Discrete Element Method

Felix Platzer, Eric Fimbinger

Abstract—In Discrete Element Method (DEM) simulations, the breakage behavior of particles can be simulated based on different principles. In the case of large, complex-shaped particles that show various breakage patterns depending on the scenario leading to the failure and often only break locally instead of fracturing completely, some of these principles do not lead to realistic results. The reason for this is that in said cases, the methods in question, such as the Particle Replacement Method (PRM) or Voronoi Fracture, replace the initial particle (that is intended to break) into several sub-particles when certain breakage criteria are reached, such as exceeding the fracture energy. That is why those methods are commonly used for the simulation of materials that fracture completely instead of breaking locally. That being the case, when simulating local failure, it is advisable to pre-build the initial particle from sub-particles that are bonded together. The dimensions of these sub-particles consequently define the minimum size of the fracture results. This structure of bonded sub-particles enables the initial particle to break at the location of the highest local loads – due to the failure of the bonds in those areas – with several sub-particle clusters being the result of the fracture, which can again also break locally. In this project, different methods for the generation and calibration of complex-shaped particle conglomerates using bonded particle modeling (BPM) to enable the ability to depict more realistic fracture behavior were evaluated based on the example of filter cake. The method that proved suitable for this purpose and which furthermore allows efficient and realistic simulation of breakage behavior of complex-shaped particles applicable to industrial-sized simulations is presented in this paper.

Keywords—Bonded particle model (BPM), DEM, filter cake, Particle breakage, Particle fracture.

I. Introduction

DURING many industrial processes, the particles of bulk solids with various material properties and shapes are broken down into smaller fragments, which can further break again. In many cases, this breakage is undesirable, such as the degradation of sinter during conveyance of this bulk material to the blast furnace since too fine-grained particles impede a sufficient gas flow in the furnace [1], when exceeding the maximum load-bearing capacity of building materials, such as concrete, in the construction industry [2] [3], or the damage of supporting rock structures in mining engineering [4] [5]. However, there are also many areas in which material breakage is desired, such as relating to crushing, grinding or drilling rock in the areas of mining engineering and mineral processing [4] [6] [7], handling cemented sands or heavily overconsolidated soils [8], or the cutting and threshing of agriculture products during harvesting [9] [10], only to name a few.

Irrespective of whether the material failure is desired or not, using a well-calibrated and for the process that is intended to be depicted suitable numerical simulation, the breakage of a particle can be replicated very accurately in a simulation environment. This is particularly suitable for optimizing processes in a very time- and cost-efficient manner.

The project on which this paper is based investigates the fracture behavior of filter cakes [11], generated from a filter press in the shape of a relatively flat plate, during the conveying process using conveyor belts and a chute system. During the material transfer, the complex-shaped cake plates break abruptly and locally, which is to be simulated by means of a numerical simulation method (specifically the Discrete Element Method (DEM), correspondingly, as a particle-based system is present). Due to the rather low moisture content of the filter cake, this material setup exhibits brittle material failure after a relatively small initial elastic deformation. The goal of this paper is to depict the macroscopic breakage behavior of sample plates made of this type of filter cake material in a sufficient way, providing a way to numerically predict if and where such filter cake plates fracture in process-like situations. It is furthermore said that this approach is thus not intended to replicate the exact microscopic crack propagation in a single sample, as the focus is set on depicting effects from a bulk-oriented perspective. To depict filter cake breakage, the Discrete Element Method (DEM) [12], also called Distinct Element Method [13], is ideal since in many other methods the simulated material is considered as a homogeneous continuum, whereas in DEM it is represented as discrete and inhomogeneous [14], which is required for the case in question. In DEM software, the change of motion and position of particles between discrete timesteps is computed based on the forces and torques acting on said particles using the laws of motion. These forces can be divided into general forces ($F_{General}$), forces due to gravity or force fields, and contact forces ($F_{Contact}$) resulting from interactions of a particle with other particles or system components.

$$F_{Particle} = F_{Contact} + F_{General} \tag{1}$$

Furthermore, the contact force is divided into master contact force F_{Master} and slave force F_{Slave}, which will be superimposed by means of superposition.

$$F_{Contact} = F_{Master} + F_{Slave} \tag{2}$$

F. Platzer is with Chair of Mining Engineering and Mineral Economics-Conveying Technology and Design Methods,Montanuniversität Leoben (University of Leoben), FranzJosef-Strasse 18, 8700 Leoben, Austria, e-mail: felix.platzer@unileoben.ac.at

E. Fimbinger is with Chair of Chair of Mineral Processing, Montanuniversität Leoben (University of Leoben), FranzJosef-Strasse 18, 8700 Leoben, Austria, e-mail: eric.fimbinger@unileoben.ac.at

Manuscript received June 3, 2022

The master contact force corresponds to the sum of forces resulting from the basic contact model, or master contact model, in the normal and tangential direction of particles in contact (cf. [15]). Additionally added slave contact models can be added to represent various physical phenomena, such as cohesion acting between moist particles, or in this case: a physical connection of particles making up a continuum.

In this paper, the DEM-based Multiphysics simulation software ThreeParticle/CAE by BECKER3D [16] was used for simulation.

II. Bonded Particle Model

Of the various methods that can be used in DEM to represent particle breakage, the Bonded Particle Model (BPM) [8] [14] [17] is ideally suited to represent local material failure under preceding deformation. In this method, two discrete particles are joined together with a virtual connection called a bond, also known as a joint or bonding. This element has neither mass nor volume but exerts loads (forces/torques) on the two particles joined by a bond as they deviate from their original relative positions. If several sub-particles are bonded together to form a cluster, any complex particle shape, from this point on called the parent particle, can be represented as such a bonded particle network. A bond connection in its initial state and under deformation is depicted in Figure 1.

In Three Particle, bonds are implemented as a slave contact model and can transmit tensile, compressive, and shear forces, as well as torque and bending moments. The forces are calculated in a local bond coordinate system with the x-axis corresponding to the bond axis, as seen in Figure 1. For this reason, all quantities required for the calculation of the bond forces and moments, such as the strain Γ, the curvature of the bonds κ, the translational as well as relative rotational velocity v_{12} resp. Ω_{rel}, are expressed in local bond coordinates.

The reaction forces $F_{s,ax}$ are calculated from the Youngs Modulus E_b, the cross-section A_b, the shear coefficient α_s, and the shear modulus of the bond G_b, according to Equation 3.

$$F_{s,ax} = \begin{bmatrix} E_b A_b & & \\ & \alpha_s G_b \ A_b & \\ & & \alpha_s G_b \ A_b \end{bmatrix} \Gamma \tag{3}$$

The viscous damping force F_d is then

$$F_d = d \ v_{12} \tag{4}$$

with the damping coefficient d. The torque acting between the bonded particles $T_{b,t}$ is calculated from

$$T_{b,t} = \begin{bmatrix} 2G_b \ J_b & & \\ & E_b J_b & \\ & & E_b J_b \end{bmatrix} \kappa \tag{5}$$

where J_b corresponds to the second moment of inertia of the beam cross-section. The damping torque T_d corresponds to

$$T_d = \begin{bmatrix} \sqrt{\frac{2G_b \ J_b}{l_0} I_r} & & \\ & \sqrt{\frac{E_b J_b}{l_0} I_r} & \\ & & \sqrt{\frac{E_b J_b}{l_0} I_r} \end{bmatrix} \Omega_{rel} \tag{6}$$

with the initial bond length l_0 and the reduced moment of inertia I_r, which is calculated from the individual moments of inertia of the particles connected with the bond according to

$$I_r = \frac{I_1 I_2}{I_1 + I_2} \tag{7}$$

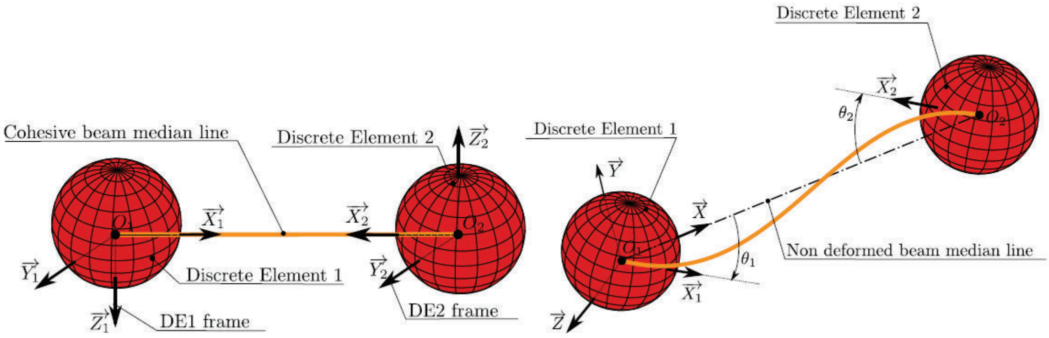
For the calculation of all above-mentioned quantities, a circular bond cross-section is assumed. According to the beam theory, the stresses can be calculated on the basis of the strain and curvature of the bond element

$$\sigma_{s,ax} = \begin{bmatrix} E_b & & \\ & G_b & \\ & & G_b \end{bmatrix} \Gamma \tag{8}$$

$$\sigma_{t,b} \quad = \begin{bmatrix} G_b & & \\ & E_b & \\ & & E_b \end{bmatrix} \kappa r_b \tag{9}$$

where r_b corresponds to the radius of the bond.

The equivalent stress is calculated using the von Mises yield criterion (cf. Equation 10). It should be noted that due to the material being much more sensitive to tensile stress, the



(a) Relaxing state

(b) Loading state

Fig. 1: Bond element connecting two particles in its initial state (left) and in the loaded state (right) [18]

normal stress component is only considered under tension and not under compression.

$$\sigma_v = \sqrt{{\sigma_{axial}}^2 + 3{\tau_{shear}}^2} \tag{10}$$

When a critical equivalent stress $\sigma_{v,krit}$ is reached, the bond is deleted, and the bonded particles experience no additional reaction forces based on this slave contact model henceforth. Due to the forces and moments being expressed in a local bond coordinate system, a final transformation into global coordinates is required.

III. Sample generation

The sub-particles are modeled as spheres, which allows a combination of suitable simulation results with a computationally efficient simulation, and also a fast generation of the complex-shaped parent particles. The generation process follows a simple scheme: A three-dimensional volume is filled with sub-particles, which are then bonded together. This procedure can easily be simulated in a DEM software, if the required shapes of the parent particles are simple, as do the calibration samples in this paper as well as the filter cake plates generated for future simulations. Although this method of generating breakable parent particles may appear to be the easiest at first glance, it quickly reaches its limits if more complex geometries are required. Furthermore, the computational effort for simulating the filling process increases exponentially with increasing particle number, as usual for DEM simulations.

Another generation method, which requires a one-time preparation effort, is the use of a filling algorithm, of which several already exist, both for arbitrarily shaped sub-particles [19] as well as for spherical ones. In this project, a filling algorithm was implemented in which three adjacent spherical sub-particles are initially placed in a seed within an arbitrary geometry given by a triangulated surface mesh, following the placement of additional sub-particles as close as possible to the already generated particles following a desired particle size distribution (PSD) [20]. As a result, the volume of the parent particle is filled with adjacent particles starting from the seed until the to be added sub-particles collide with the surface mesh, as can be exemplarily seen in in Figure 2. During the filling process, each newly placed particle is directly bonded to its immediate neighbor resulting in a bonding network shown in Figure 2e. The generated parent-particle consists of 41,788 sub-particles and 235 220 bonds.

Depending on the material, care must be taken not to introduce any preferred crack paths [21] into the parent particle during generation, which can be easily controlled when using a filling algorithm. In addition to a much shorter generation time of the parent particle by means of a filling algorithm, compared with the simulation of the filling process, an algorithm is also characterized by the fact that the computation time increases significantly slower with increasing particle number.

IV. Calibration test

To obtain results from the simulation of a process in a suitable form, meaning the depiction of the correct macro-

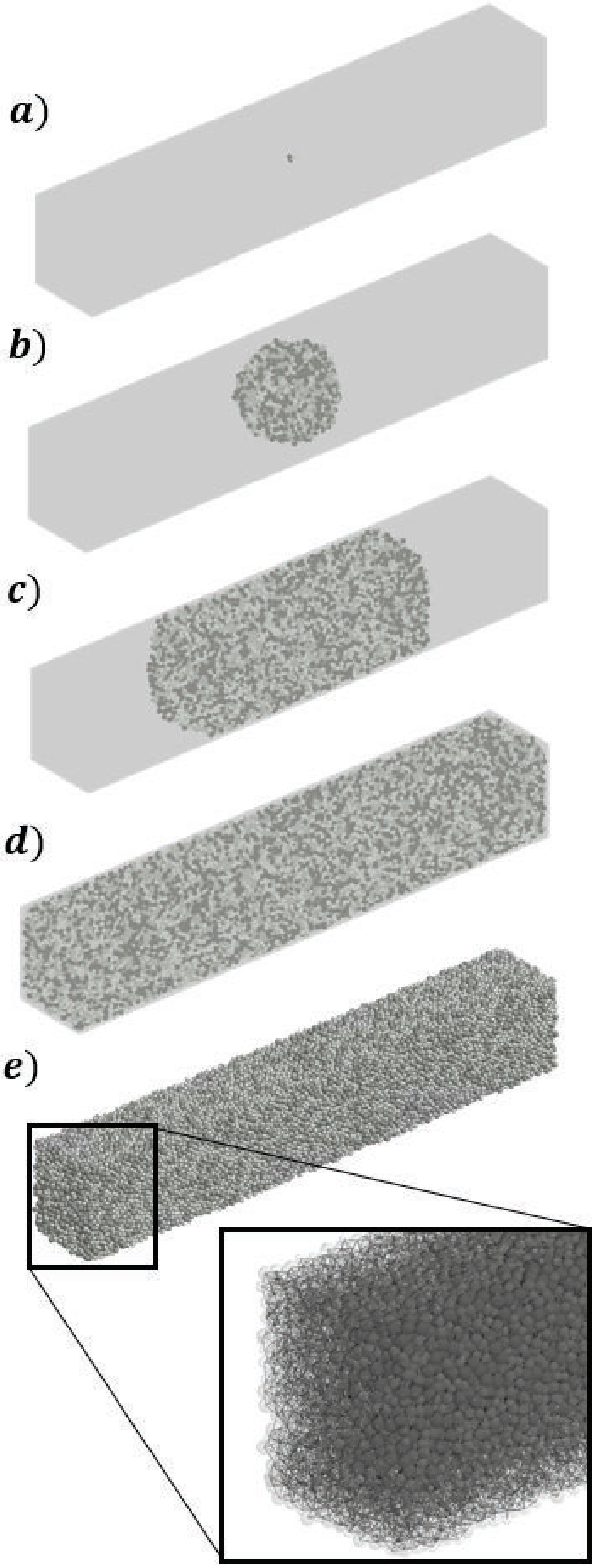


Fig. 2: Visualization of the sample generation using a filling algorithm starting with a seed within the desired geometry specified by an STL file (a), the filling process (b-d) and the finished parent particle within the DEM software including a closeup of the bonds connecting the sub-particles (e)

scopic behaviour,the simulation micro parameters must be sufficiently calibrated. For this purpose, suitable calibration tests, i.e. reflecting the loads prevailing in the final process, must be selected. The calibration of typical simulation parameters, such as particle density or friction coefficients between particles, is not further discussed in this paper, since values in these contexts can be determined with commonly-known, standardized tests, such as an angle of repose test (cf. [22]).

For the calibration of rock and rock-like materials, whereby rock-like materials are understood to be materials that exhibit the same failure criteria as rock and, above all, brittle material behavior [23], numerous standardized tests already exist. These tests are used to determine characteristic material parameters, such as tensile or compressive strength. Due to the tensile strength of the filter cake being significantly lower than the compressive strength, as is common for rock-like materials, and the parent particles being initialized in thin plates in the final simulation, resulting the parent particle to most likely fail due to bending, i.e. due to tensile stress in the edge fiber, standardized tests that measure tensile strength are chosen to calibrate the bond parameters. Thus, the four-point bending flexural test is considered since it is not only used to measure tensile strength but also leads to material failure due to bending. The three-point bending flexural test is not considered in detail due to its dependence of the results on the specimen shape [24].

The procedure of the four-point bending flexural test, as well as the sample dimensions, are specified in various international standards for testing rock or cement -based products (e.g. EN 13161, ASTM C880-89). Since, in this application, the height of the specimen is specified by the filter press, a slightly modified sample geometry and position of the force application are selected. In the laboratory test, the sample is loaded with a force F divided between two loading points, as can be seen in Figure 3a, to subject the specimen to a constant bending moment and no shear forces between those points. The force F is increased continuously to ensure a constant deformation rate v of 0.0002 m/s until the material failure occurs under the maximum force F_{max} at a deformation of $w_{s,max}$ depicted in Figure 3b, while measuring the applied forces as well as the deformation.

The length L_S of the test sample was set at 180 mm, the width b_S at 30 mm, the sample thickness h_S given by the filter press is 35 mm and the distance between the loading points d_S at 80 mm. The resulting mean µ of the measured values as well as the standard deviation σ of several tests are shown in Figure 4.

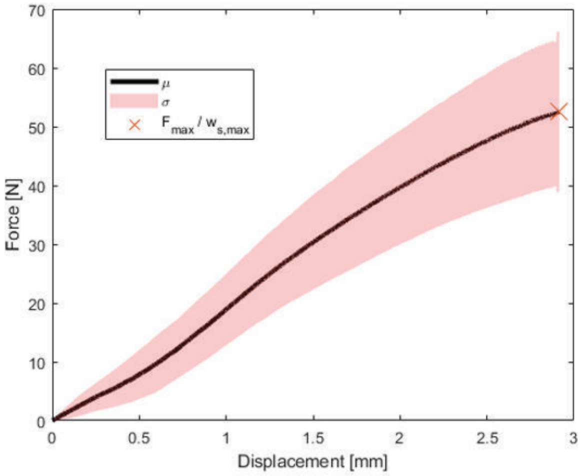


Fig. 4: Mean and standard deviation of the measured force over the displacement

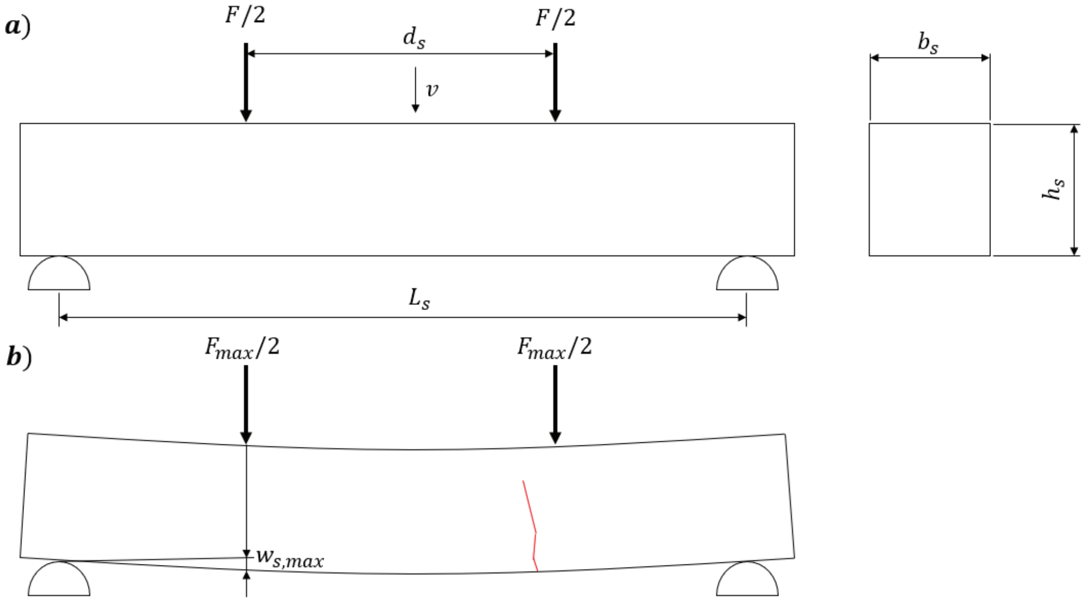


Fig. 3: Schematics of the four-point bending flexural test, showing the beginning of the experiment (a) and the breaking of the test sample (b)

To calibrate the parameters of the bonds, the laboratory test is replicated in the simulation environment with the parent - particle being deformed at the same rate until failure occurs. A comparison of the laboratory test with the simulation results is shown in Figure 5, where in this case, the broken bonds are highlighted instead of deleted to better visualize the fracture of the simulated test sample.

Subsequently, the deformation path of the simulated sample as well as the required force are compared with the values measured in the laboratory tests, see Figure 6.

When comparing the results of the laboratory tests and the simulation, it can be seen that the failure occurs at the same force and deformation. Due to averaging several laboratory tests, the curves deviate from each other. However, since these local deviations are marginal and the material limit state is depicted accurately this is considered negligible.

V. Simulation parameters

To ensure an efficient calibration of the simulation model, the parameters with the most significant influence on the simulation results are determined. These parameters are both general simulation parameters, such as the timestep and

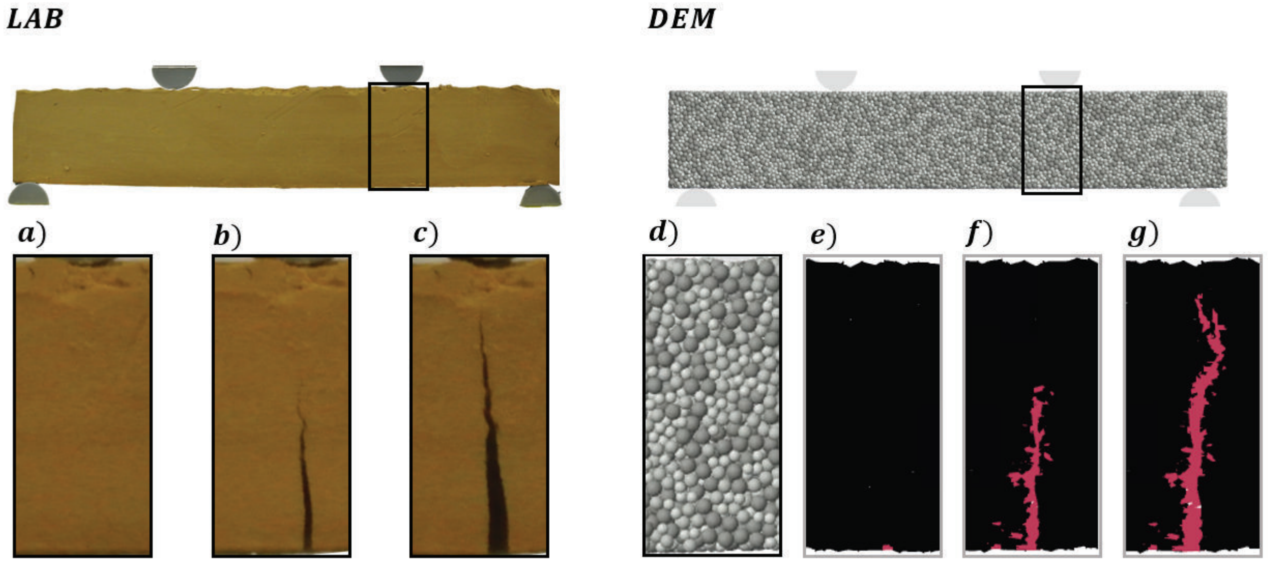


Fig. 5: Visual comparison of the material failure in the calibration test (LAB) and the simulation (DEM) under the maximum force F_{max}. Closeup of test sample (a) and crack propagation (b and c), as well as closeup of the simulated parent-pacticle with the sub-particles visible at first (d), then depicting a cross-section with only the bonds visualized (e) and finally the crack propagation (f and g) with the broken bonds highlighted.

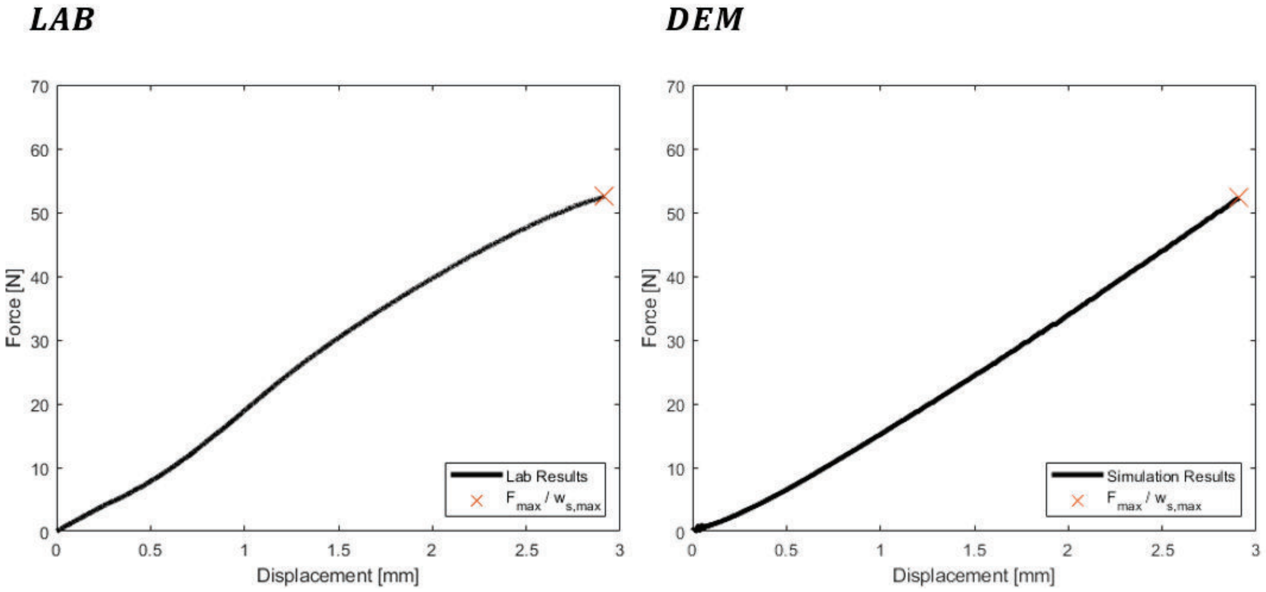


Fig. 6: Comparison of the measured data from the laboratory tests (LAB) and the simulation environment (DEM)

particle size, as well as model-specific micro-parameters.

1) Timestep: The timestep (Δt) has a great influence on a DEM simulation since its size affects the computational efficiency as well as the stability and accuracy of the simulation. For this reason, it should be chosen as large as possible but not to exceed a critical value at which the simulation tends to become unstable. Influenced by several different simulation parameters, this critical value also depends on the type of interaction.

For particle-particle interactions or particle interacting with system components, about 20% of the Rayleigh timestep $t_{Rayleigh}$ [25] is usually used to determine a suitable timestep, which corresponds to

$$t_{Rayleigh} = \frac{\pi \;\; R \;\; \sqrt{\frac{\rho}{G}}}{0.1631 \;\; \upsilon \; + \; 0.8766} \tag{11}$$

with the particle Radius R, density ρ, shear modulus G and Poisson's ratio υ, all being constant for all particles except the Radius, resulting in the critical timestep being that of the smallest particle Radius.

For bonded structures the critical value is calculated from the critical vibration frequency of the particles connected with massless bonds [26], following

$$t_{Bond,crit.} = \; 0.17\sqrt{\frac{m}{K}} \tag{12}$$

in three-dimensional space, with the particle mass m and bond stiffness K. This value is determined for the smallest, and therefore the lightest, particles bonded together with the highest bond stiffness. The timestep used in the simulation is the smallest critical timestep calculated from bonded and non-bonded contacts.

$$\Delta t = min(t_{Rayleigh}; t_{Bond,crit.}) \tag{13}$$

2) Bond Stiffness: The stiffness of the bonds corresponds to the macroscopic Youngs Modulus, which can be measured by means of local instrumentation during the calibration tests. In this case, the material's Youngs Modulus E is calculated from the beam deformation w_s at a loading point under a load F according to the elastic beam theory

$$w_s = \; \frac{F}{48 \;\; E \;\; I} \;\; (L_S - d_S)^2 \, (L_S + 2 \; d_S) \tag{14}$$

with the second moment of area I of the sample cross-section. By transforming this equation and applying the average values resulting from the experiments performed (the values for deformation and force at which material failure occurs, $w_{s,max}$ and F_{max}), the macroscopic Young's modulus is calculated, resulting in 13.2 MPa.

It is most convenient to initially deactivate the ability of the bonds to break when calibrating the bond's stiffness. The lower the stiffness of the bonds, the less force must be applied to deform the parent-particle to achieve the desired state of deformation. If a lower force at the deformation at which the fracture occurs in the laboratory test is measured in the simulation with the bond stiffness set as the macroscopic Youngs Modulus, the stiffness is continuously increased until the exact laboratory values are reproduced, respectively decreased if a higher force is measured.

3) Critical stress: In order to reproduce the results of the laboratory tests, Tthe maximum bond tensile stress at the desired fracture point is evaluated from the calibration simulation, which is then checked, and fine-tuned, if necessary, in subsequent simulations. In addition to this, the results can be checked for plausibility by calculating the tensile stress due to the bending moment according to Equation 15.

$$\sigma_b = \; \frac{3F(L_S - d_S)}{2 \; b_S {h_S}^2} \tag{15}$$

4) Sub-particle size: Although the size of the sub-particles is not considered a classic simulation parameter, it has a considerable influence on the results of the calibration simulation. That is why the relation between the smallest distance within the sample geometry L to the average sub- particle diameter d must be taken into account when establishing the PSD used in the simulation. When testing the compressive strength of Rock the American Society for Testing and Materials (ASTM) recommends a ratio of L to the maximum grain size d_{max} of 10, while the International Society for Rock Mechanics (ISRM) suggests this ratio to be at least 20. When applied to DEM modeling, a ratio of L/d of 25 is recommended to keep the coefficient of variation of most model parameters under 2% [28].

VI. Conclusion

Complex-shaped particles capable of breaking are encountered in a wide range of technical processes, which are consequently often to be analyzed via numerical simulations using

Property	Describtion	Value	
ρ	Particle density	3 720	[kg/m³]
G	Particle Shear modulus (reduced, as typicall; cf [27])	10	[MPa]
μ	Particle interaction friction coefficient	0.67	[-]
d_{min}	Minimum Particle diameter	0.7	[mm]
d_{max}/ d_{min}	Ratio of maximum to minimum particle diameter	1.43	[-]
L/d	Ratio of characteristic length of the parent particle to median particle diameter	35	[-]
Eb	Bond Youngs Modulus	12.8	[MPa]
$\sigma_{v,krit}$	Critical equivalent bond stress	0.24	[MPa]

TABLE I: Summary oft the DEM simulation parameter

the DEM. In this paper, a computationally-efficient filling algorithm was used to generate arbitrarily shaped parent-particles composed of sub-particles, which are furthermore connected with deformable beams, termed bonds. An efficient way to choose the general simulation parameters as well as to calibrate the Bonded Particle Model parameters was shown by means of a four-point bending flexural test. This allows an easy way of generating, calibrating, and consequently simulating complex-shaped, deformable, and furthermore breakable bodies representing a brittle, rock-like material behavior in an efficient and suitable form to depict relatively large amounts of bulk media containing such complex types of DEM particles.

VII. Outlook

In order to simulate industrial-sized processes with several differently shaped parent-particles, further aspects have to be taken into consideration. Besides maintaining a realistic mass and volume flow during the breakage process, the optimization in regard to computational efficiency, as well as the consideration of the dynamic behavior of the parent-particles, the possibility to automatically detect different sub-particle clusters resulting from the breakage of bonded particle structures is of great interest.

References

[1] Michael Denzel and Michael Prenner. “Minimierung des Sinterzerfalls mittels DEM”. In: *BHM Berg- und Hüttenmännische Monatshefte* 166.2 (2021), pp. 76–81. ISSN: 0005-8912. DOI: 10.1007/s00501-021-01081-7.

[2] Eugenio Oñate et al. “A local constitutive model for the discrete element method. Application to geomaterials and concrete”. In: *Computational Particle Mechanics* 2.2 (2015), pp. 139–160. ISSN: 2196-4378. DOI: 10.1007/s40571-015-0044-9.

[3] Peter Domone and Marios Soutsos, eds. *Construction materials: Their nature and behaviour*. 5. ed. Boca Raton: CRC Press, Taylor & Francis Group, 2018. ISBN: 1315164590.

[4] Zong-Xian Zhang. *Rock Mechanics Related to Mining Engineering*. Helsinki, Finland, October 11–12, 2017.

[5] Peixian Li, Lili Yan, and Dehua Yao. “Study of Tunnel Damage Caused by Underground Mining Deformation: Calculation, Analysis, and Reinforcement”. In: *Advances in Civil Engineering* 2019 (2019), pp. 1–18. ISSN: 1687-8086. DOI: 10.1155/2019/4865161.

[6] Johannes Quist and Carl Magnus Evertsson. “Cone crusher modelling and simulation using DEM”. In: *Minerals Engineering* 85 (2016), pp. 92–105. ISSN: 08926875. DOI: 10.1016/j.mineng.2015.11.004.

[7] R. A. Bearman, C. A. Briggs, and T. Kojovic. “The applications of rock mechanics parameters to the prediction of comminution behaviour”. In: *Minerals Engineering* 10.3 (1997), pp. 255–264. ISSN: 08926875. DOI: 10.1016/S0892-6875(97)00002-2.

[8] Martin Obermayr et al. “A bonded-particle model for cemented sand”. In: *Computers and Geotechnics* 49 (2013), pp. 299–313. ISSN: 0266352X. DOI: 10.1016/j.compgeo.2012.09.001.

[9] Petre Miu. *Combine Harvesters: Theory, modeling, and design*. Boca Raton: CRC Press, 2015. ISBN: 9780429152931. DOI: 10.1201/b18852. URL: https://www.taylorfrancis.com/books/9781482282375.

[10] Qirui Wang, Hanping Mao, and Qinglin Li. “Modelling and simulation of the grain threshing process based on the discrete element method”. In: *Computers and Electronics in Agriculture* 178 (2020), p. 105790. ISSN: 01681699. DOI: 10.1016/j.compag.2020.105790.

[11] Todd Wisdom, Mike Jacobs, and James Chaponnel. “GeoWasteTM – continuous comingled tailings for large-scale mines”. In: *Proceedings of the 21st International Seminar on Paste and Thickened Tailings*. Proceedings of the International Seminar on Paste and Thickened Tailings. Australian Centre for Geomechanics, Perth, 2018, pp. 465–472. DOI: 10.36487/ACG_rep/1805_38_Wisdom.

[12] P. A. Cundall and O. D. L. Strack. “Discussion: A discrete numerical model for granular assemblies”. In: *Géotechnique* 30.3 (1980), pp. 331–336. ISSN: 0016-8505. DOI: 10.1680/geot.1980.30.3.331.

[13] John A. Hudson, ed. *Comprehensive rock engineering: Principles, practice & projects*. 1. ed. Oxford [u.a.]: Pergamon Press, 1993. ISBN: 9780080406152.

[14] Nicholas J. Brown, Jian-Fei Chen, and Jin Y. Ooi. “A bond model for DEM simulation of cementitious materials and deformable structures”. In: *Granular Matter* 16.3 (2014), pp. 299–311. ISSN: 1434-5021. DOI: 10.1007/s10035-014-0494-4.

[15] Peter Wriggers and B. Avci. “Discrete Element Methods: Basics and Applications in Engineering”. In: *Modeling in Engineering Using Innovative Numerical Methods for Solids and Fluids*. Ed. by Riva, Laura de Lorenzis, and Alexander Düster. Vol. 599. CISM International Centre for Mechanical Sciences. Cham: Springer International Publishing, 2020, pp. 1–30. ISBN: 978-3-030-37517-1. DOI: 10.1007/978-3-030-37518-8_1.

[16] *ThreeParticle/CAE*. URL: http://becker3d.com/.

[17] D. O. Potyondy and P. A. Cundall. “A bonded-particle model for rock”. In: *International Journal of Rock Mechanics and Mining Sciences* 41.8 (2004), pp. 1329–1364. ISSN: 13651609. DOI: 10.1016/j.ijrmms.2004.09.011.

[18] Damien André et al. “Discrete element method to simulate continuous material by using the cohesive beam model”. In: *Computer Methods in Applied Mechanics and Engineering* 213-216 (2012), pp. 113–125. ISSN: 00457825. DOI: 10.1016/j.cma.2011.12.002.

[19] Y. Ma et al. “Packing Irregular Objects in 3D Space via Hybrid Optimization”. In: *Computer Graphics Forum* 37.5 (2018), pp. 49–59. ISSN: 01677055. DOI: 10.1111/cgf.13490.

[20] Elias Lozano et al. “An efficient algorithm to generate random sphere packs in arbitrary domains”. In: *Computers & Mathematics with Applications* 71.8 (2016), pp. 1586–1601. ISSN: 08981221. DOI: 10.1016/j.camwa.2016.02.032.

[21] H. A. Carmona et al. “Fragmentation processes in impact of spheres”. In: *Physical Review E* 77.5 Pt 1 (2008), p. 051302. ISSN: 1539-3755. DOI: 10.1103/PhysRevE.77.051302.

[22] C. J. Coetzee. “Review: Calibration of the discrete element method”. In: *Powder Technology* 310 (2017), pp. 104–142. ISSN: 00325910. DOI: 10.1016/j.powtec.2017.01.015.

[23] Jinjin Ge and Ying Xu. “A Method for Making Transparent Hard Rock-Like Material and Its Application”. In: *Advances in Materials Science and Engineering* 2019 (2019), pp. 1–14. ISSN: 1687-8434. DOI: 10.1155/2019/1274171.

[24] A. Coviello, R. Lagioia, and R. Nova. “On the Measurement of the Tensile Strength of Soft Rocks”. In: *Rock Mechanics and Rock Engineering* 38.4 (2005), pp. 251–273. ISSN: 0723-2632. DOI: 10.1007/s00603-005-0054-7.

[25] Rayleigh. “On Waves Propagated along the Plane Surface of an Elastic Solid”. In: *Proceedings of the London Mathematical Society* s1-17.1 (1885), pp. 4–11. ISSN: 00246115. DOI: 10.1112/plms/s1-17.1.4.

[26] Catherine O’Sullivan and Jonathan D. Bray. “Selecting a suitable time step for discrete element simulations that use the central difference time integration scheme”. In: *Engineering Computations* 21.2/3/4 (2004), pp. 278–303. ISSN: 0264-4401. DOI: 10.1108/02644400410519794.

[27] Stef Lommen, Dingena Schott, and Gabriel Lodewijks. “DEM speedup: Stiffness effects on behavior of bulk material”. In: *Particuology* 12 (2014), pp. 107–112. ISSN: 16742001. DOI: 10.1016/j.partic.2013.03.006.

[28] Xiaobin Ding et al. “Effect of Model Scale and Particle Size Distribution on PFC3D Simulation Results”. In: *Rock Mechanics and Rock Engineering* 47.6 (2014), pp. 2139–2156. ISSN: 0723-2632. DOI: 10.1007/s00603-013-0533-1.

Determination of Mechanical and Contact Parameters for DEM Simulation of Direct Reduced Iron

A. Hossein Madadi-Najafabadi, Masoud Nasiri

Abdolhossein Madadi-Najafabadi is with the Mobarakeh Steel Company, Iran (e-mail: a.madadi@msc.ir).

Abstract

The discrete element method is a powerful technique for numerical modeling the flow of granular materials such as direct reduced iron. It would enable us to study processes and equipment related to the production and handling of the material. However, the characteristics and properties of the granules have to be adjusted precisely to achieve reliable results in a DEM simulation. The main properties for DEM simulation are size distribution, density, Young's modulus, Poisson's ratio and the contact coefficients of restitution, rolling friction and sliding friction. In the present paper, the mentioned properties are determined for DEM simulation of DRI pellets. A reliable DEM simulation would contribute to optimizing the handling system of DRIs in an iron-making plant. Among the mentioned properties, Young's modulus is the most important parameter, which is usually hard to get for particulate solids. Here, an especial method is utilized to precisely determine this parameter for DRI.

Keywords: Direct reduced iron; Discrete element method; Young’s modulus; Contact parameters

1. Introduction

Today, direct reduction (DR) process can be considered as a well-developed iron-making route in the world [1]. The increasing need for direct reduced iron (DRI) in the steel production industries which use electric arc furnaces, made it as an important product in the world. The total global production of DRI in 2018 has been more than 100 million tones [2].

DRIs vulnerability to successive collisions and fine generation is the Achilles Heel of such companies. During DR process, transportation and buffer storage, DRI pellets are subjected to successive pellet-pellet and pellet-wall collisions. Depending on the impact conditions (i.e. angle, velocity and force) and the pellets characteristics, various kinds of mechanical damages (attrition, fragmentation or abrasion) may occur to the pellets [3-5].

Therefore, the industries have been trying to improve mechanical strength of iron ore pellets (IOP) to reduce DRI pellets damages; IOPs are the primary material of DRIs. Several researchers have studied on the breakage and abrasive wear of IOPs [4-6]; however, less has been done on DRIs damages.

The most important work in this relation was done by Boechat et al. recently [7]. They studied the effect of DR process on the mechanical characteristics of IOP in order to estimate their extent of damages inside DR furnace. They used the model of Boechat et al. [8] to estimate the amount of DRI's mass loss during collisions. The mentioned model determine the extent of surface damage of DRIs and IOPs based on the reference collision energy and mass-specific energy.

Application of the numerical, analytical and experimental methods to analysis DRIs abrasive damage, would result in improvement of transportation and storage systems in the iron making plans. Such improvements would decrease DRIs mass loss and fine generation during DRIs handling from DR plant to electric arc furnaces.

From the analytical or experimental analyses, some predictive models for DRIs damages are obtainable. Such models estimate the extent of damage as a function of material characteristics (hardness, fracture toughness, mass specific energy and etc.) and collisions parameters (energy, force and velocity). The collisions parameters are obtained by simulation approaches such as discrete element method (DEM).

DEM is one of the most powerful methods for simulation of granular materials such as IOP and DRI. It is capable of accurately calculating the parameters of particle-particle and particle-wall collisions. However, the accuracy of DEM results depends directly on the precision of the material parameters which are set as the input variables of DEM simulation.

Barrios et al. [9] determined mechanical and contact parameters of IOPs using single-particle tests to ensure reliable DEM simulation of this granular material.

Here, the mechanical and contact parameters of DRI pellets produced in Mobarakeh steel company (MSC) are determined using specific methods to provide accurate and reliable DEM simulation of this granular material.

2. Material

A 20 kg sample of DRI pellets was selected from transfer line of MSC (after screen) for the following tests. The chemical compositions, density, porosity (according to ASTM C20) and crushing strength (according to …, using RB 1000) of the selected DRIs which are routine tests of the laboratory are shown in Table 1. Among the mentioned properties density is necessary for DEM simulation and the rest are reported just to specify the quality and conditions of the tested DRIs in this study.

Table 1. Chemical, physical and mechanical properties of the tested DRIs.

Property		Value
Chemical (%)	Total iron	87.39
	Degree of metallization	93.74
	Metallic iron	82.01
Porosity (%)		45.5
Crushing strength (kg)	Mean value	137.4
	Standard deviation	50.45
	Min / Max	63 / 230

3. Methods

3.1. Size distribution

Size distribution of sampled DRIs is identified based on ASTM E11. The DRIs were sampled after screen, therefore they are in the range of 9 to 16 mm.

Size distribution in DEM simulation should be set based on the mean values of the intervals. Considering low strength of DRIs, their size distribution is changed during transportation. Thus, the DRIs size distribution should be determined in every project specifically.

3.2. Poisson's ratio

A brief review of Hertz-Mindlin's contact force shows that the value of Young's modulus has a considerable effect on the magnitude of calculated forces. But Poisson's ratio does not have a significant effect on the contact forces.

Bruno et al. considered Poisson's ratio of particulate solids is independent of porosity and is equal to that of the solid domain [10]. Given that, more than 90 percent of chemical composition of DRI is pure iron, its Poisson's ratio is estimated to be in the range of 0.21 to 0.3, based on Poisson's ratio of different types of iron.

Considering arbitrary values for Young's modulus, mass, radius and coefficient of restitution for DRIs, the change of Poison's ratio from 0.21 to 0.3 results in about 5 percent variation in calculated forces.

3.3. Young's modulus

Gustafsson et al. [11] used Bruno et al. [10] method for calculation of Young's modulus for iron ore pellet, but considering the ultra-porous structure of DRI, his equation is not usable for DRI.

Tavares and King [12] introduced a specific set-up using ultrafast load cell to measure particle stiffness during breakage. Young's and shear modules of particulate solids are calculated based on the particle stiffness. However, DRI pellets do not exhibit similar behavior in failure due to the ultra-porous and brittle structure. Noting that, the value of Young's modulus is desired for DEM simulation of DRIs, it should be determined as an elastic constant value without occurrence of breakage.

Here, a simple method based on the contact duration in a collision is used to obtain Young's modulus of DRI. The method achieves Young's modulus through elastic collision of spheres without occurrence of breakage. The results show very good repeatability and acceptable accuracy of the presented method.

For colliding spheres, Patricio [13] defined collision duration as the time in which deformation goes from 0 to the maximum value and back to 0 again. He calculated the mentioned time as Eq. 1, based on Hertz contact theory.

$$T_c = 3.21\left(\frac{\mu^2}{e^2 r v}\right)^{1/5} \qquad (1)$$

Where v is relative velocity and μ, r and e are reduced mass, reduced radius and reduced elastic constant, respectively, and are obtained using Eqs. (2) to (4) in terms of mass (m), radius (R), Poisson's ratio (σ) and Young's modulus (E) of the spheres 1 and 2.

$$\mu = \frac{m_1 m_2}{m_1+m_2} \tag{2}$$

$$r = \frac{R_1 R_2}{R_1+R_2} \tag{3}$$

$$e = \frac{4}{3}\left(\frac{1-\sigma_1^2}{E_1} + \frac{1-\sigma_2^2}{E_2}\right)^{-1} \tag{4}$$

Obviously, for given parameters of mass, radius, Poisson's ratio and Young's modulus, Eq. 1 is a function of collision duration in terms of collision velocity.
Minamoto and Kawamura [14] obtained contact duration of colliding spheres based on the pulse produced by the closing of an electric circuit as shown in Fig. 1.

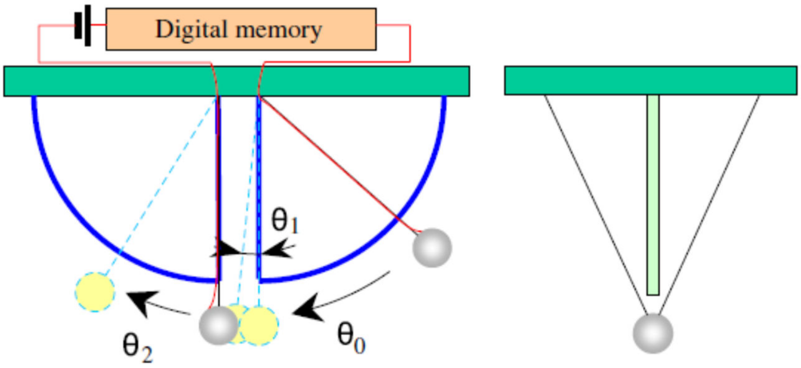


Fig. 1. Pendulum impact setup of Minamoto and Kawamura [14]

Based on the experimental results, they achieved an approximation expression in the form of Eq. 5 for SUJ2 steel.

$$T_c = AV_i^B \tag{5}$$

Where V_i is collision velocity and A and B are obtained using regression technique.
Considering electrical conductivity of DRI, the experimental method of Minamoto and Kawamura is applicable for DRI. Thus, their experimental approach was conducted to obtain the coefficients A and B for collision of a steel ball to DRI pellets. The characteristics of the steel ball are presented in Table 1.

Table 1. Characteristics of the steel ball.

Mass (g)	Radius (mm)	Poisson's ratio	Young's modulus (GPa)
8.5	6.3	0.3	208

An ultra-thin copper wire was precisely soldered to the steel ball and three DRI pellets and the collision duration was measured using a specific microcontroller. In order to calibrate the set-up, time duration for collision of two steel balls (characterized in Table 1) was measured in different

collision velocities. The time duration for steel balls collisions were also theoretically obtained using Eq. 1. Considering the theoretical values as reference, a correction factor was obtained for the experiment.
For the main experiments, a DRI pellet of specified weight was initially in rest and the steel ball was released from a specified pendulum angle, θ_0. Before each test, the radius of DRI pellet in the contact point was measured using radius gauge. The experiment was repeated with five specified values of θ_0 to test five collision velocities of 0.5, 0.7, 1, 1.5 and 2 m/s for each DRI. Each collision test was repeated several times to find five perfectly normal collisions and the mean value of collision duration of the five collisions was considered.

3.4. Coefficient of restitution.

Minamoto and Kawamura [14] utilized the described pendulum impact setup in section 3.1, and the equations 6 to 9 in order to measure the coefficient of restitution of steel spheres. They measured the maximum angles of the each pendulum after impact (θ1, θ2) from the video taken by a digital camera.

$$e = \frac{V_2 - V_1}{V_i} = \frac{\sqrt{1-\cos\theta_2} - \sqrt{1-\cos\theta_1}}{\sqrt{1-\cos\theta_0}} \qquad (9)$$

Where, Vi, V1 and V2 indicate the initial impact speed, post-impact speed of the impacting sphere, post-impact speed of the target sphere, respectively.
Here, the same setup was used for collisions of DRI pellets to each other and to the steel ball. The maximum angles θ1 and θ2 were measures based on the videos captured by digital camera in 200 frames per second. The experiment was repeated with ten DRI pellets at different collision velocities from 0.5 to 5 m/s.

3.5. Coefficient of rolling friction

In DEM simulation of the granular materials with irregular sphere shape there is always the problem of shape mismatch. Researchers used to solve this problem in two ways:

- multi-sphere approximation of the granules,
- adjustment of the coefficient of rolling friction.

Barrios et al. [9] studied both solutions for IOPs on the basis of rolling angle of single IOPs on a flat surface. They considered two shape parameters of aspect ratio and sphericity to characterize the shape of IOPs. They validated the results based on the slump experiment (angle of repose) and a specific tumbler test. They predicted the coefficient of rolling friction for both spherical and overlapping models. Their overlapping sphere model which was simulated in EDEM is shown in Fig. 5.

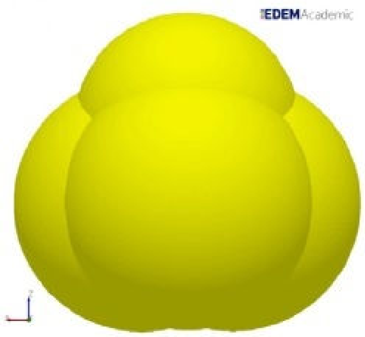


Fig. 5. Overlapping sphere model of Barrios et al. […]

Due to the exact similarity of the shape of DRIs to IOPs, the results provided by Barrios et al. for IOPs seem to be applicable for DRIs. In present work, the parameter of aspect ratio of 30 DRI pellets is measured to compare to that of IOPs (reported by Barrios et al.).

3.6. Coefficient of sliding friction

Madadi-Najafabadi et al. [6] introduced an innovative approach for measuring coefficient of sliding friction of IOP. They used a lath machine equipped with a dynamometer and two pellet holding devices as shown in Fig. 6 to measure normal and tangential components of contact force in a tangential collision of two IOPs. They determined the coefficient of sliding friction by dividing the tangential component by the normal component.

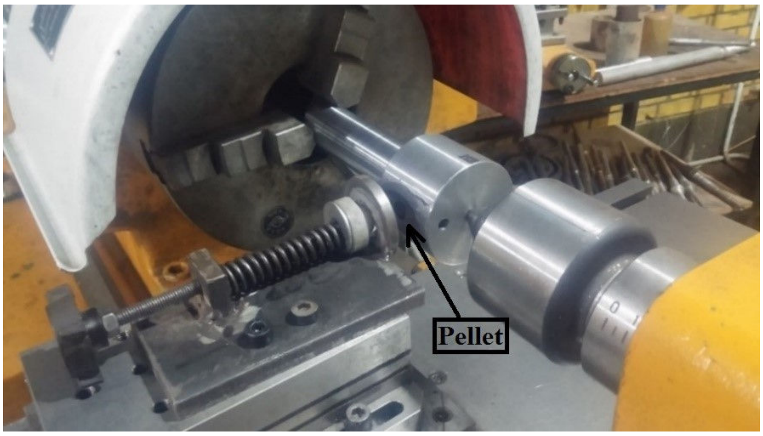


Fig. 6. The test setup introduced by Madadi et al. for measurement of the coefficient of sliding friction of granular materials [6].

Here, the same method is utilized to obtain coefficient of sliding friction in DRI- DRI and DRI-steel contacts. The experiment was repeated for ten DRI couples in collision velocities of 2, 3.5 and 5 m/s.

4. Results

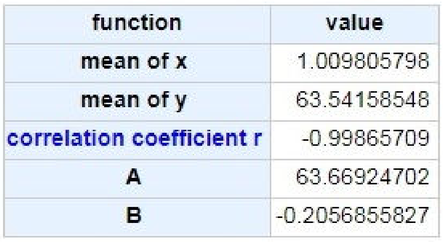
4.1. Size distribution

Size distribution of the sampled DRIs is shown in Table 2. The mean value and percentage of each interval is specified for DEM simulation. To increase reliability of DEM simulation, it is recommended to use more sieves and segregate DRIs to more size batches.

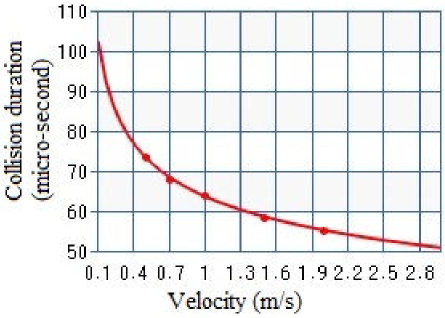
Table 2. Size distribution of sampled DRIs

4.2. Young’s modulus

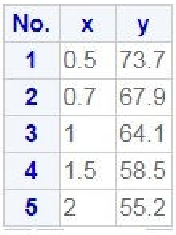
For three tested DRIs mentioned in section 2.3, the experimental results, fitted curve and obtained coefficients (A and B) are calculated using regression technique and are presented in Figs. 2 to 4.



function	value
mean of x	1.009805798
mean of y	63.54158548
correlation coefficient r	-0.99865709
A	63.66924702
B	-0.2056855827

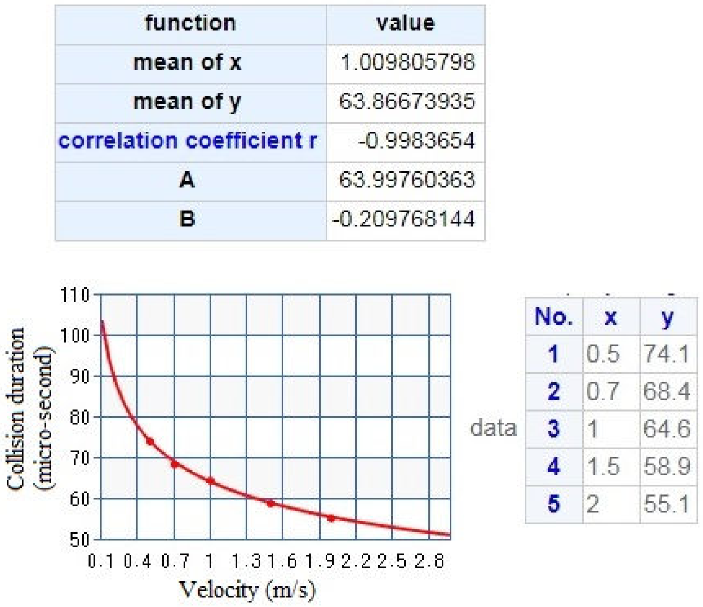


data



No.	x	y
1	0.5	73.7
2	0.7	67.9
3	1	64.1
4	1.5	58.5
5	2	55.2

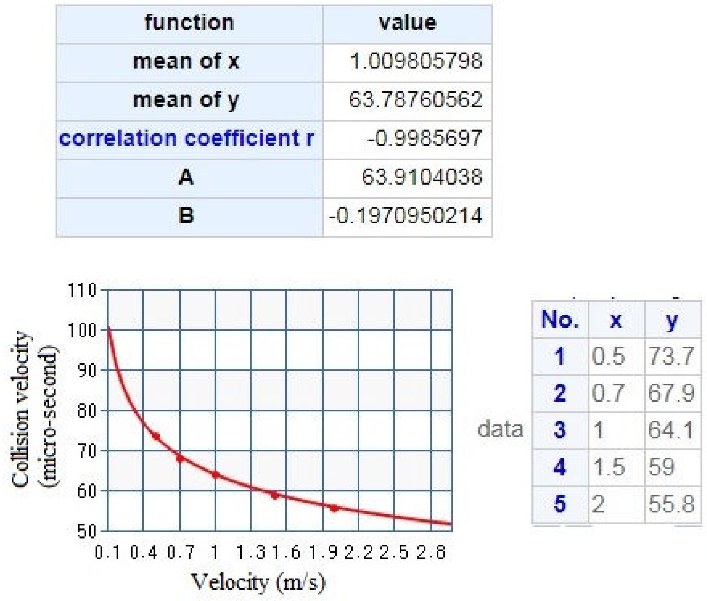
Fig. 2. Experimental results, fitted curve and the values of A and B for DRI1.



function	value
mean of x	1.009805798
mean of y	63.86673935
correlation coefficient r	-0.9983654
A	63.99760363
B	-0.209768144

No.	x	y
1	0.5	74.1
2	0.7	68.4
3	1	64.6
4	1.5	58.9
5	2	55.1

Fig. 3. Experimental results, fitted curve and the values of A and B for DRI2.



function	value
mean of x	1.009805798
mean of y	63.78760562
correlation coefficient r	-0.9985697
A	63.9104038
B	-0.1970950214

No.	x	y
1	0.5	73.7
2	0.7	67.9
3	1	64.1
4	1.5	59
5	2	55.8

Fig. 4. Experimental results, fitted curve and the values of A and B for DRI3.

The power coefficients (B) are close to -0.2, which shows very good agreement with Eq. 1. The coefficient A of the obtained functions are also close. Averaging A values and considering B = -0.2, a final function for contact duration of collision of the steel ball and the DRIs is acquired as $T_c = 63.86 \times 10^{-6} V_i^{-0.2}$ in second.

Considering $A = 63.86 \times 10^{-6}$, from the Eqs. 1 and 5, reduced elastic constant (e) of tested DRIs are calculated. Mass, radius (in contact point) and obtained values of reduced elastic constant of tested DRIs are presented in Table 2.

Table 2. Obtained values of e for the tested DRI pellets.

DRI	Mass (g)	Radius (mm)	e (GPa)
1	4.1	6.5	27.29
2	4.5	7	26.36
3	4.6	7	28.93

From the results, an average of e = 27.53 GPa is obtained for the tested DRIs. Therefore, considering e = 27.53 GPa and $0.21 < \nu_{\mathrm{DRI}} < 0.3$, using Eq. 4, Young's modulus of the tested DRI pellets is calculated to be $20.65 < E < 21.7$ GPa. Finally, the mean value of 21.175 GPa is estimated as Young's modulus of tested DRI pellets.

4.3. Contact parameters

The obtained values for coefficients of restitution and sliding friction of DRI-DRI and DRI-steel contacts, are presented in Table 3.

Table 3. coefficients of restitution and sliding friction of DRI-DRI and DRI-steel contacts

Parameters	DRI-DRI	DRI-steel
Coefficient of sliding friction	0.45	0.46
Coefficient of restitution	0.51	0.43

Table 4 presents the aspect ratio of tested DRIs in present study and IOPs reported by Barrios et al.

Table 3. Mean value and standard deviation of aspect ratio of tested DRIs (present study) and IOPs (reported by Barrios et al. [9]) and the estimated values of coefficient of rolling friction of IOPs by Barrios et al.

IOP	Measurement	0.9 (0.055)[a]
	Model (overlapping)	0.876 (0.056)[a]
	Model (sphere)	1.00[a]
DRI	Measurement	0.96 (0.042)
[a] Barrios et al. [9]		

Clearly, the obtained mean value for aspect ratio of DRIs is very close to that of the value measured by Barrios et al. for IOPs. Therefore, the values of coefficient of rolling friction in contact of IOPs to each other and to steel, estimated by Barrios et al., are applicable for DEM simulation of DRIs. The mentioned values are presented in Table 4.

Table 4. Coefficient of rolling friction for IOP-IOP and IOP-steel contact, estimated by Barrios et al. from single pellet calibration tests and simulations [9].

Model	Contact type	
	Pellet - pellet	Pellet – steel
Overlapping spheres	0.02	0.01
Sphere	0.21	0.25

5. Conclusion

Reliability of DEM simulation of a granular material depends strongly to the accuracy and precision of the parameters and characteristics which are set in DEM simulation. There are different techniques to estimate the mentioned parameters each have their own advantages and disadvantages.

Here, the main parameters and characteristics of DRI for DEM simulation were estimated using single particle methods. The methods were choose considering the properties and conditions of DRI based on the previous studies.

Among them, Young’s modulus which is the most important parameter for DEM simulation was estimated on the basis of non-destructive collision tests. It was obtained according to the contact duration in normal collision of the pellets. The coefficient of restitution was measured using the same setup based on the pellets velocity before and after collision.

The coefficient of sliding friction was estimated using a previously introduced method in which the contact force components (normal and tangential) are measured during a single collision.

In order to solve the irregular shape problem, two simulation models and the appropriate coefficient of rolling friction were mentioned. The models and related coefficients were previously introduced for IOP by Barrios et al.; considering the similarity of the aspect ratio of DRIs and IOPs, the mentioned solutions are applicable for DRIs.

For density and size distribution chart, common laboratory measuring techniques were utilized and for non-significant parameter of Poisson’s ratio a range was determined.

The general characteristics of tested DRIs including chemical composition, porosity and crushing strength were reported in section 2 to let other researchers compare their own DRIs to the tested DRIs in present work. For DRIs with significantly different porosity, chemical composition and crushing strength, it is strongly recommended to determine Young’s modulus and coefficient of restitution using the method described in this paper. However, the other parameters are allowed to be used for variety of DRIs from different plants.

References

Synthesis, Anti-Inflammartory Activity of 3-Amino 5-Methoxyl-2-Methyl Quinazolin-4(3H)-one an Amino-6-Methoxyl-2-Methyl of 4H–Benzo[d] [1,3]–Oxazine–4–one

Osarumwense Peter Osarodion

***Abstract*—**

Introduction: Quinazolinone derivatives represent one of the most active classes of compounds possessing a wide spectrum of biological activity. They are widely used in pharmaceuticals and agrochemicals.

Methods: The condensation of 2-amino-methyl-5-dimethoxybenzoate with acetic anhydride yielded the cyclic compound 2-methyl-5-substituted-1,3-benzo-oxazine-4-one which further produced a novel 2,3-disubstituted quinazolin-4-ones via the reaction with hydrazine hydrate. The compounds synthesized were unequivocally confirmed by means of Infrared, Nuclear Magnetic Resonance (1H and 13C), Gas chromatography-mass spectrophotometer and elemental analysis. The synthesized compounds were screened and evaluated pharmacologically for their in-vivo anti-inflammatory activity and the paw volume of each rat was measured before 1 h and after 3 h of carrageenan treatment with the help of a plethysmometer.

Results: Compound 1 had anti-inflammatory activity of 89.03% and 88.03% at 20 mg/kg and 10 mg/kg respectively, while compound 2 had anti-inflammatory activity of 94.79% and 90.30% at 20 mg/kg and 10 mg/kg respectively.

Discussion: Compound 1 displayed a singlet signal at: δ 3.78 attributed to methoxy group and singlet at δ 3.68 which was due to methyl group. Also, 1H NMR spectrum of compound 2 showed a characteristic signal at δ 2.56 (singlet) corresponding to methyl group and duplet at: δ 3.90 for methoxy group. For the IR spectra, Compound 1 was characterized by absence of v NH2 and presence of v C-O stretch in 1101cm-1 region of the compound. Compound 2 showed the highest anti-inflammatory activity at 20 mg.kg of 94.79% compared to compound 1 and indomethacin. These compounds synthesized had a higher anti-inflammatory activity than indomethacin which is a standard anti-inflammatory drug.

Conclusion: Compound 2 had a higher anti-inflammatory activity than Compound 1. The compounds synthesized had a higher anti-inflammatory activity than Indomethacin, a standard anti-inflammatory drug.

***Keywords*—** anti-inflammatory activity, quinazoline-4(3H)-One, 6-methoxyl 2-methyl 4H–benzo[d] [1,3]–oxazine–4–One, nucleo phile, synthesis, 3-Amino 6-methoxyl -2-methyl.

Osarumwense Peter Osarodion is with the Ondo State University of Science and Technology, Nigeria (e-mail: osarodionpet@yahoo.com).

How Safe Do Dutch Students Handle Their Password Security?

A. Ali, D. Avé, E. De Leeuw, T. Kester, H. Alers, R. Van Der Kleij

***Abstract*—With the internet becoming increasingly more relevant in our lives, online security and password security are becoming more important every day. From ordering a pizza to organizing finances and bank accounts, most online activities require an account and therefore a password. This research is conducted to get a better view on how Dutch students handle their password security and if there are any significant differences between the education levels MBO, HBO and WO towards password security. To determine whether this is the case, an online survey was put together. Questions within the survey were based on the COM-B model of behavior (Capability, opportunity and motivation). The objective for this study was to investigate how capable students are to behave safely online, if students get the opportunity to create safe passwords and if students are motivated to create safe passwords. The results from the online survey demonstrated that most students did not meet the required score to be considered sufficient on multiple of the tested areas. This means that most students on the three mentioned education levels are not handling their password security safe enough. Only one significant difference was found with the analyses of the collected data. This difference is the knowledge score between WO and HBO students, with WO students scoring significantly better.**

***Keywords*—** behavior, cybercrime, password security, students

I. Introduction

THE past few years, the number of cybercrimes has increased dramatically. In 2019, 13% of the population above 12 years old reported being a victim of cybercrime, with the largest percentage between the ages of 15 and 24. In 2020, about 127% more reports were made, that is about 16.5% of the population [1]. Those are only the percentage of people who were actual victims of cybercrimes. More than 60% of the Dutch population has had to do with phishing emails or WhatsApp fraud texts. A big part of the reason why the cybercrime numbers are increasing is because of the growing online world. Especially in the year 2020, online activity became more important due to the COVID-19 pandemic [2].

Cybercrimes are crimes that are committed with ICT-resources like a computer, smartphone or tablet. There are six forms of cybercrime that are common: hacking, ransomware, phishing, identity fraud, DDOS-attacks and human failure. Phishing is the most common (39%), then comes ransomware (20%) and DDOS-attacks (14%) [1]. Hacking involves shutting down or abusing your website or network. Phishing involves getting sent false emails to collect your personal information. Identity fraud is the abusing of your personal information by using your credit card for instance. DDOS-attacks means the server gets overloaded, human mistakes is just people accidently leaking data. The causes of cybercrimes include human failure (31%), cyber-attacks (44%) and technical failures (25%). [3]

Being a victim of any of the cybercrimes explained in the paragraph above can lead to financial damage, but also diverse forms of emotional damage. Therefor it is important for people to know when and how to prevent yourself from becoming a victim of cybercrimes. However, a study commissioned by the national government shows that the Dutch are not always aware of digital risks or choose to ignore them. There is no need among the Dutch to improve their own online security. 60% of the Dutch are already taking action, but the 40% that don't, have no need for it either. Moreover, most Dutch people do not want to pay for their online security. It also appears that despite the fact that young people on average have more knowledge about the digital world and its security, they still have a less alert attitude to it. [4]

It is clear now that Dutch people are poorly aware of the risks even though the chances of becoming a victim is growing increasingly. It seems that the younger generations take more risks despite being more aware of the menace than older generations. Different studies have been conducted about password security and people their behavior to understand it and sometimes even improve it. For instance a study commissioned by R. van der Kleij, S. van 't Hoff-de Goede, S. van de Weijer and R. Leukefeldt about the behavior of Dutch people on the internet [4]. This research showed that people think they are behaving a lot safer than they actually are. It also showed that knowledge is not necessarily the problem and stated that the behavior could be improved by creating opportunities. Another research that was conducted about password behavior took place in the United States of America [5]. This research was focused on students and faculty staff of a university in America and showed that 80% uses the same set of passwords for everything, while 60% uses the same password but with a slight modification. Both researches concluded that the overall behavior of people online is not safe.

Rick van der Kleij, a professor at the Hague university who also conducted the research about password behavior in the Netherlands, now wants to research Dutch students specifically. The interests in Dutch students comes from the fact that they were not concluded in any other researches about password behavior and the fact that they are part of the generation that takes the most risks online.

Eline de Leeuw is with the The Hague University of Applied Sciences, Netherlands (e-mail: elinedleeuw@gmail.com).

This research is conducted to learn more about the password security of Dutch students. The main research question is:

How safe do Dutch students handle their password security?

II. Methodology

The data gathering method that is used for this research is a survey. With a survey, questions can be asked about all the COM-B aspects (capability, opportunity and motivation). Also, knowledge questions in combination with self-assessment questions are asked to compare if students know as much about password security as they claim to know. The questions are asked in a Likert-scale based form. [6],[7]

Structure:
To get a clear answer to this question, a few sub-questions are determined, the sub-questions are:

1. How knowledged are Dutch students with different levels of education about password security?
2. Do Dutch students with different levels of education have the opportunity to be secure?
3. How motivated are Dutch students with different levels of education about password security?
4. What is the difference in behavior towards password security between different levels of education?

The population for this research are Dutch students in the Netherlands. Within this group a distinction is made between MBO students, HBO students and WO students from the Netherlands. This research includes national and international Dutch students but excludes nationals who study outside the Netherlands.

The sample size is calculated with a confidence level of 80% and a margin of error of 5,5%. With a population of 1.330.685 Dutch students, the sample size needs to be 138 students. Table 1 shows the percentages of each education level.

The sample size is calculated with the following equation:

$$n = 2(\frac{(z * \sigma)}{m})$$

$$where:$$

$$n = Samplesize$$

$$Confidence\ level\ at\ 95\%, z = 1{,}96$$

$$\sigma = Standard\ Deviation = for\ 3\ groups, 33\% = 0{,}33$$

$$m = Margin\ of\ Error = 5{,}5\% = 0{,}055$$

$$n = (\frac{(1{,}96 * 0{,}33)}{0{,}055})^2 = 138{,}3 = 138\ Students$$

	Number of students	Percentage
MBO Students	509797	38%
HBO Students	489383	37%
University/ WO Students	331505	25%
Total	1330685	100%

Table 1 Total number of students in The Netherlands per education level [8]

A. Com-B

Previous research about cyber behavior of Dutch residents [4] concludes that the differences between reported behavior and the actual behavior is significantly bigger than expected. When approaching behavior in this research, the COM-B model of behavior is leading [9].

The COM-B model focusses on three factors that influence behavior, capability, opportunity and motivation. These three factors are capable of changing behavior.

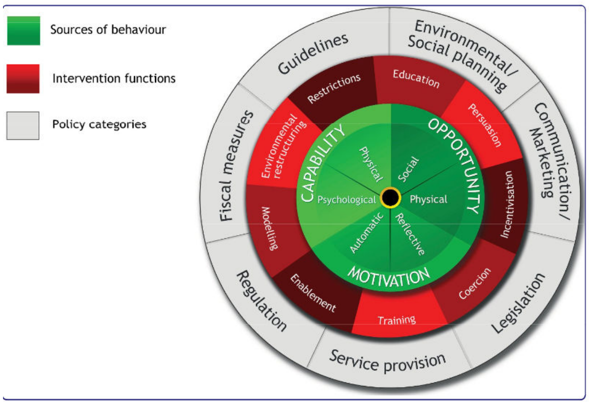


Fig. 1 COM-B Model of behavior [10]

Capability refers to a person's psychological and physical ability to participate in an activity. Opportunity refers to external factors that make a behavior possible. Motivation refers to how motivated a person is about the expected behavior.

At least one of these components need to change for a behavior change to occur. [11] In short, this model states that when capability and opportunity are present, motivation will increase and therefor positive behavior will be seen more frequently. The COM-B model also provides information about different kinds of interventions and what interventions can effectively change capability, opportunity or motivation. [6] The COM-B model is used a lot over the years to explain behavior but using it in the context of cybersecurity is quite new.

Other than capability, opportunity and motivation, knowledge also plays a big part in the behavior towards password security. Someone who knows more about online and password security tends to be more aware about the dangers, this leads to them creating stronger passwords and behaving more secure overall [12].

B. Survey Design

The sample for the research consists of 138 Dutch students. To reach this mount of respondents, different tactics were used. Students have been approached via social media (Instagram, WhatsApp and Linked In). Also, an email list of different schools was created and an email was sent to all these schools. The email asked if the schools wanted to collaborate and share the survey with their students. To get as many respondents as possible, student associations were asked to share the survey with their members as well.

C. Survey Questions

Here is given an overview of the survey questions.

- *Demographics* - First the respondents educational level, field of study, age and gender were asked.
- *Safety Knowledge* – The respondents safety knowledge was tested using multiple choice questions.
- *Password Behavior* – The respondents were asked several behavior questions; whether they write down their password or if they share their passwords.
- *Password Creation Strategy* – Respondents were asked if they had to make an account now, would they reuse their old password? Respondents were also asked if they made a new password what strategy the used to make the password. Options included a variety of options, from names of something or someone to phone numbers and birthdays.
- *Password Composition* – Respondents were asked to think of an account and answer the questions about its password. They had to answer if the account forces them to use a minimum and maximum of characters, if so, how many? They were asked if the password for the account had any symbols, capital or small letters or any numbers in it and if they believed that the password was safe.

Survey Administrations and analysis

The survey questions were categorized into the following categories:

- Capability
- Password knowledge
- Password handling
- Password strength
- Opportunity & motivation

A scoring system is assigned to questions. Extra weight is factored into the questions that are more important for security. The formula that is used to calculate the final score is:

$$Capability = x + y + z$$

$$x = 2(ScorePS + 13)$$

$$y = ScorePH + 35.5$$

$$z = ScoreK + 50$$

$$ScorePS = 2a + 2b + 5c + d + e$$

$$ScorePH = 3f + 2g + 5h + i + j + 3k$$

$$ScoreK = l + m + n + 3o + p + q + r$$

ScorePS represents password strength, ScorePH password handling, ScoreK knowledge and a through r represent the survey questions. A respondent scoring above 25 for password strength is considered to have adequate password strength. For password knowledge 25 and for password handling a respondent scoring above 20. For capability, a respondent scoring above 70 is considered to have adequate capability.

III. Results & analysis

Here will the results gotten from the survey be described. First, the *demographics* of the participants are detailed, followed by the findings related to *password knowledge.* Next, there is a description of the findings to *password handling, password strength* and *capability.* Finally, the findings related to *opportunity and motivation.*

A. Demographics

				Frequency	Percentage
Gender	Male	Age	16-18 years	9	6.5%
			19-21 years	28	20.3%
			22-24 years	16	11.6%
			25-27 years	5	3.6%
			27+ jaar	0	0%
			Total	58	42%
	Female	Age	16-18 years	26	18.8%
			19-21 years	34	24.6%
			22-24 years	18	13%
			25-27 years	2	1.4%
			27+ jaar	0	0%
			Total	80	58%

Table 2 Gender and age of the participants, frequency & percentage.

There is date collected from 138 participants, 56 participants are male (40.6%) and 72 participants are female (59.4%), see table 2. Since males and females are more or less balanced and the age falls in logical student age range there can be concluded that the data is representative of the population.

		Frequency	Percentage
Educati on Level	MBO	53	38%
	HBO	51	37%
	WO	34	25%

Table 3 Education level of the participants, frequency & percentage

From the 138 participants 53 participants (38%) have an educational level of MBO, 51 participants (37%) HBO and 34 participants (25%) WO as shown in table 3. Therefore, the data is representative of the population

B. Password Knowledge

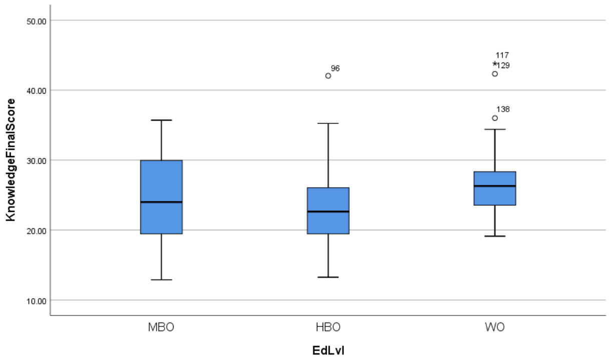


Fig. 2. Password Knowledge final scores of students in MBO, HBO and WO

Students were asked several *password knowledge* related questions. To assess whether a password knowledge, knowledge was operationalized in two ways: general digital security knowledge and an assessment on their passwords. For each of these questions a score was given. The sum of this score is the *KnowledgeFinalScore.* The highest respondents could score was 53. In Fig. 2 MBO, HBO and WO students are displayed on the X-axis and the *KnowledgeFinalScore* on the Y-axis. The first boxplot represents MBO students. The second boxplot represents HBO students and the third boxplot represents WO students. The line in the box represents the mean, the box represents the first and third quartile and the whiskers the minimum and maximum value. The circles represent the outliers.

The mean score for the password knowledge based questions between all education levels is 24.59 with HBO and MBO scoring 23.24 and 24.32 respectively. This is not enough since a respondent must score above 25 to be considered having adequate password handling. WO students with a mean score of 27.06, did score above 25 which is considered to be sufficient.

The respondents had to answer the question ”which of the following passwords is considered a strong password”. Only 22.5% of the respondents choose the right answer.

A one-way between subjects ANOVA was conducted to compare the scores of password knowledge for MBO, HBO and WO students. There was a significant difference of Password Knowledge at the $p<0.05$ level for the three groups [F(2, 135) = 4.50, p=0.013]. Post hoc comparisons using the Tukey HSD test indicated that the mean score for HBO students (M= 23.24, SD= 6.17) was significantly different than WO students (M= 27.06, SD= 5.59). However, the mean score of MBO students (M=24.32, SD= 6.17) did not significantly differ from the HBO and WO students. Taken together, these results suggest that WO students have more password knowledge than HBO students.

MBO students do not appear to have a significantly different password knowledge than HBO or WO students.

C. *Password Handling*

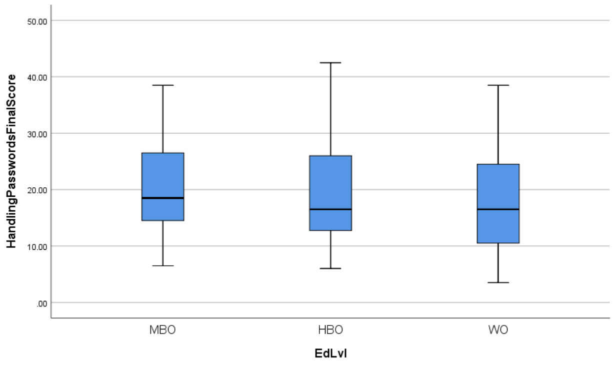


Fig. 3 Handling password final score of the students in the three education levels.

Students were asked several *password handling* related questions. For each of these questions a score was given. The sum of this score is the *HandlingPasswordsFinalScore.* The highest respondents could score was 45.5. In Fig. 3 MBO, HBO and WO students are displayed on the X-axis and the *HandlingPasswordsFinalScore* on the Y-axis. The first boxplot represents MBO students. The second boxplot represents HBO students and the third boxplot represents WO students. This Fig. uses the same style as the boxplots prior.

The mean score for the password handling based questions between all education levels is 19.8 with HBO and WO scoring 19.7 and 18.4 respectively. This is not enough since a respondent must score above 20 to be considered having adequate password handling. MBO students with a mean score of 20.8, did score above 20 which is considered to be sufficient.

A one-way between subjects ANOVA was conducted to compare the scores of password handling for MBO, HBO and WO students. There was no significant difference of Password Handling at the $p<0.05$ level for the three groups[F(2, 135) = 0.70, p=0.497]. These results suggest that MBO, HBO and WO students do not significantly differ in their password handling.

D. *Password strength*

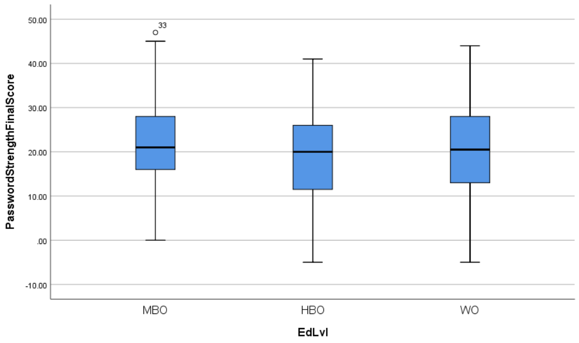


Figure 4 Password strength final score of the students in the three education levels.

Students were asked several *password strength* related questions. Password Strength represents the current behavior of respondents and are self-assessment questions. For each of these questions a score was given. The sum of this score is the *PasswordStrengthFinalScore.* The highest respondents could score was 52. In figure 4 MBO, HBO and WO students are displayed on the X-axis and the *PasswordStrengthFinalScore* on the Y-axis. The first boxplot represents MBO students. The second boxplot represents HBO students and the third boxplot represents WO students. This Figure uses the same style as the boxplots prior.

The mean score for the questions about password strength between all education levels is 20.6 with MBO scoring the highest with 22.5. This is still not enough since a respondent must score above 25 to be considered having adequate password strength. Of respondents 60,9% consider their password safe and 61,6% considered their behavior as safe behavior. The answers ‘strongly agree’ and ‘agree’ were included in this calculation.

A one-way between subjects ANOVA was conducted to compare the scores of password strength for MBO, HBO and WO students. There was no significant difference of password strength at the $p<0.05$ level for the three groups [F(2, 135) = 1.51, p=0.224]. These results suggest that MBO, HBO and WO students do not significantly differ in the strength of their passwords.

E. Capability

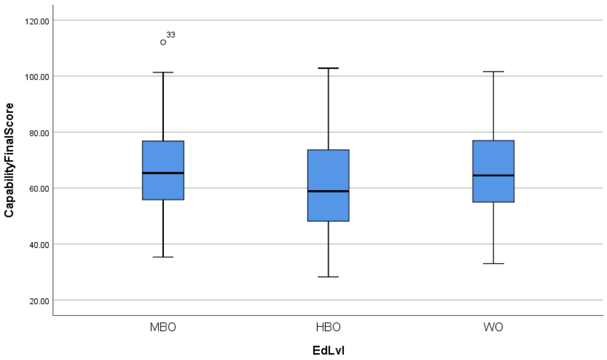


Fig. 5 Capability final score of the students in the three education levels.

Combining the scores of the password strength, password knowledge & password handling, password capability is calculated. This score is the *CapabilityFinalScore.* The highest respondents could score was 150.5. In Fig. 5 MBO, HBO and WO students are displayed on the X-axis and the *CapabilityFinalScore* on the Y-axis. The first boxplot represents MBO students. The second boxplot represents HBO students and the third boxplot represents WO students. This Fig. uses the same style as the boxplots prior.

The mean score for the capability based questions between all education levels is 64.9 with MBO scoring the highest with 67.6. This is still not enough since a respondent must score above 70 to be considered having adequate capability.

A one-way between subjects ANOVA was conducted to compare the scores of capability for MBO, HBO and WO students. There was no significant difference of capability at the p<0.05 level for the three groups [F(2, 135) = 1.51, p=0.225]. These results suggest that MBO, HBO and WO students do not significantly differ in their capability.

F. Opportunity

		Education Level					
		MBO		HBO		WO	
		Frequency	Percentage	Frequency	Percentage	Frequency	Percentage
Opportunity Minimum Characters	Yes	38	28%	33	24%	29	21%
	No	7	5%	3	2%	0	0%
	I don't know	8	6%	15	11%	5	4%

Table 4 Opportunity question about minimum characters, frequency in the three education levels.

To measure opportunity the students were asked to take a particular password in mind and answer the question '*With the creation of your account, was a minimal number of password characters required?'.*

A one-way between subjects ANOVA was conducted to compare the opportunity of whether there is a minimum amount of characters required when creating a password for MBO, HBO and WO students. There was no significant difference of opportunity at the p<0.05 level for the three groups [F(2, 135) = 2.09, p=0.128]. These results suggest that MBO, HBO and WO students do not significantly differ in whether there is a minimum number of characters required when they are creating a password.

		Education Level					
		MBO		HBO		WO	
		Frequency	Percentage	Frequency	Percentage	Frequency	Percentage
Opportunity Maximum Characters	Yes	14	10%	8	6%	2	1%
	No	19	14%	18	13%	13	9%
	I don't know	20	14%	25	18%	19	14%

Table 5 Opportunity question about maximum characters, frequency in the three education levels.

A one-way between subjects ANOVA was conducted to compare the opportunity of whether there is a maximum amount of characters required when creating a password for MBO, HBO and WO students. There was no significant difference of opportunity at the p<0.05 level for the three groups [F(2, 135) = 3.00, p=0.053]. These results suggest that MBO, HBO and WO students do not significantly differ in whether there is a maximum number of characters required when they are creating a password.

G. Motivation

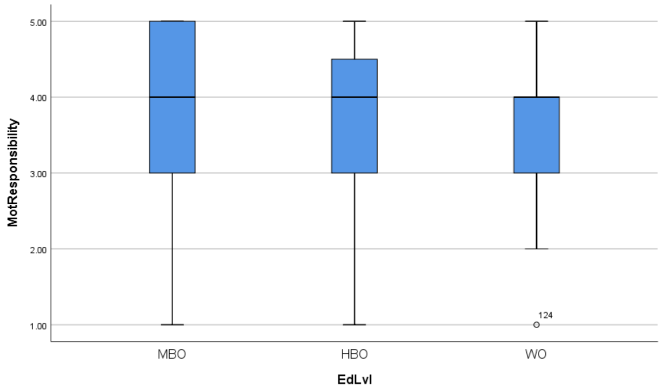


Fig. 6 MBO, HBO and WO students opinion on responsibility for their own digital safety

Students were asked whether they thought they were responsible for their own digital safety. MotResponsibility is a Likert question where the values range from Strongly Disagree (1) to Strongly Agree (5). In Fig. 6 MBO, HBO and WO students are displayed on the X-axis and the *MotResponsibility* on the Y-axis. The first boxplot represents MBO students. The second boxplot represents HBO students and the third boxplot represents WO students. Fig. 6 uses the same style as the boxplots prior.

A one-way between subjects ANOVA was conducted to compare the opinion on responsibility for your own digital safety for MBO, HBO and WO students. There was no significant difference of opinion at the p<0.05 level for the three groups [F(2, 135) = 2.21, p=0.113]. These results suggest that MBO, HBO and WO students do not significantly differ in their opinion on whether they are responsible for their own digital safety. 82,6% of the respondents say that they think having a strong password. For this percentage the answers ‘Strongly agree’ and ‘agree’ are included.

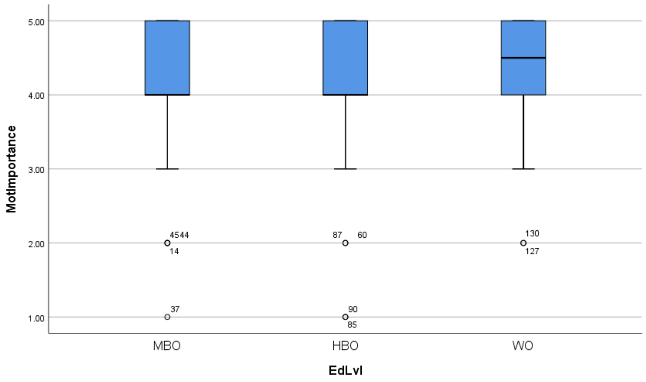


Fig. 7 Students opinion about importance of a strong password in all three the education levels.

Students were asked whether they thought having a strong password is important for their digital safety. MotImportance is a Likert question where the values range from Strongly Disagree (1) to Strongly Agree (5). In Fig. x MBO, HBO and WO students are displayed on the X-axis and the *MotImportance* on the Y-axis. The first boxplot represents MBO students. The second boxplot represents HBO students and the third boxplot represents WO students. This Fig. uses the same style as the boxplots prior.

A one-way between subjects ANOVA was conducted to compare the opinion on having a strong password is important for your digital safety for MBO, HBO and WO students. There was no significant difference of opinion at the p<0.05 level for the three groups [F(2, 135) = 0.61, p=0.546]. These results suggest that MBO, HBO and WO students do not significantly differ in their opinion on whether having a strong password is important for your digital safety. 61,6% of the respondents think it is their own responsibility to create a good password and handle it safely. For this percentage the answers ‘Strongly agree’ and ‘agree’ were included.

IV. Conclusion

Since males and females are more or less balanced and the age falls in logical student age range there is to believe the data is representative of the population. The results suggest that WO students have significantly more password knowledge than HBO students. Still, MBO students do not appear to have a significantly different password knowledge than HBO or WO students. The only group that that scored above the required 25 for password knowledge are WO students, MBO and HBO students scored insufficiently. For password strength the three groups did not meet the requirement of 25. MBO scored the highest with 22.5. There was no significant difference in password strength between the three groups.

For password handling, MBO students with a score of 20.8 were the only one to score above the required score of 20 points, HBO and WO students scored insufficiently. The results suggest that there is no significant difference in password handling for the three groups.

Out of all respondents, 73% percent answered that there was a minimum number of characters required when creating a password. However, 48% of all respondents do not know whether there is a maximum amount of characters they can use when creating a password. Taken together with the result that the three groups do not significantly differ from each other the results suggest that students do have the opportunity to be secure. MBO, HBO and WO students believe they are responsible for their own digital safety. Also they believe that having a strong password is important for their digital safety.

Remarkable is that 82,6% of the respondents say that they think having a strong password and handling it safely is important, 61,6% of the respondents think it is their own responsibility to create a good password and handle it safely. Also 60,9% considered their password safe and 61,6% considered their behavior as safe behavior, while the data shows that most respondents are not behaving responsible with their passwords.

On many of the tested areas, the mean of the students did not meet het required score needed to qualify as adequate. Therefore it is safe to say that most Dutch students do not handle their password security safely, the data shows that this is the case with all different levels of education. Except for knowledge between WO and HBO students, no significant differences were found.

V. Discussion

A. Interpretation of results

After analyzing the data a small pattern showed, majority of the participant think that having a strong password and handling it safely is important. But the data shows that most respondents are not behaving responsible with their passwords. The expectations where that there would be a significant difference in password security among different levels of education. There was also expected that the students would handle their password security with little care which the data represents.

The result of this research are more and less similar as those of a Previous study[4],about the security behavior of Dutch citizen. The result of that study showed that Dutch citizen think their online behavior is safe but the data showed that it’s not.

B. Implications

This research was the first one to be focused on Dutch students. So there are no studies to compare to regarding the different levels of study. However a previous study[4], conducted by Rick van der Kleij and colleagues, showed that Dutch people think their behavior online is safe, but actually is not. These results are similar to the results. They also added that knowledge was not the problem for Dutch citizens. This was different from this study, since this study showed that students over all do not have enough knowledge about password security. In the Dutch study ‘Ons cybergedrag is veel onveiliger dan we zelf denken’ conducted by Rick van der Kleij and colleagues [4] there was concluded that people weren´t motivated enough or did not have enough opportunity to create a safe password. In this research however, was shown that motivation on itself was not the problem. Respondents indicated they thought password security is very important and is their own responsibility. One thing is clear: It is safe to say that there is no harm in researching how to improve Dutch students their behavior regarding password security.

C. Limitations of the study

Participants in this study were not random from the population, since most were from social inner circles. If this was not done, it would have resulted in a low participation rate. This would be problematic since there is no way of knowing the results apply to non-participants.

There is only measured how Dutch students think about how safe they are handling their password security. It is assumed all respondents answered the questions truthfully, but there is always the possibility of people lying or simply not knowing their own behavior. With an experiment it would have been possible to measure their actual behavior and compare it to the answers they give us. Due to time restriction it was not possible. Their actual behavior is measured through questions in the survey, but it is not as accurate as with an experiment, which could be interesting to do in future experiments. Nevertheless, future research is needed to study actual password behavior objectively.

During the research a few difficulties arose. For instance, when trying to contact different Universities and colleges to help us reach students, they often responded with no due to their privacy rules. This made finding enough respondents, especially MBO students, difficult. The goal of reaching 272 students was not achieved and the sample size needed to be changed, this was done by changing the margin of error from 4% to 5.5%.

Acknowledgment

Recognized by The Hague University of Applied Sciences. This research was funded by the above-mentioned researchers.

References

[1] Hoe zit het met cybercrime?. (2021). Retrieved 18 September 2021, from https://longreads.cbs.nl/nederland-in-cijfers-2020/hoe-zit-het-met-cybercrime/

[2] Criminaliteit 2020: minder inbraak, meer cybercrime (2021). Retrieved 18 september 2021, from https://www.politie.nl/nieuws/2021/januari/15/00-criminaliteit-2020-minder-inbraak-meer-cybercrime.html

[3] Welke soorten cybercrime zijn er?- Zicht adviseurs. (2021). Retrieved 18 September 2021, from https://www.zichtadviseurs.nl/zakelijk/bedrijf/cybercrime-verzekering/vormen

[4] ons cybergedrag is veel onveiliger dan we zelf denken (2020), Retrieved 18 september 2021, from https://www.researchgate.net/publication/343273232_Ons_cy

bergedrag_is_veel_onveiliger_dan_we_zelf_denken_Implicaties_voor_effectief_beinvloedingsbeleid_door_de_overheid

[5] Shay, R., Komanduri, S., Gage Kelley, P., Giovanni Leon, P., Mazurek, M. L., Bauer, L., Christin, N., & Faith Cranor, L. (n.d.). Encountering Stronger Password Requirements: User Attitudes and Behaviors. Retrieved September 8, 2021, from www.incommonfederation.org

[6] An application and empirical test of the Capability Opportunity Motivation-Behaviour model to data leakage prevention in financial organizations (2020), Retrieved 17 September 2021, from https://www.sciencedirect.com/science/article/pii/S0167404820302431

[7] How to analyze Likert and other rating scale data (2015), from https://www.sciencedirect.com/science/article/pii/S1877129715200196

[8] CBS. (2021) StatLine - Leerlingen, deelnemers en studenten; onderwijssoort, woonregio. Retrieved September 19, 2021, from https://opendata.cbs.nl/statline/#/CBS/nl/dataset/71450ned/table?fromstatweb

[9] Michie, S., Stralen, van, M., West, R. (2011). The behaviour change wheel: A new method for characterising and designing behaviour change interventions

[10] Michie, S., Stralen, van, M., West, R. (2011). The behaviour change wheel: A new method for characterising and designing behaviour change interventions.

[11] https://thedecisionlab.com/reference-guide/organizational-behavior/the-com-b-model-for-behavior-change/

[12] https://thedecisionlab.com/reference-guide/organizational-behavior/the-com-b-model-for-behavior-change/

A Survey of Recognizing of Daily Living Activities in Multi-User Smart Home Environments

Kulsoom S. Bughio, Naeem K. Janjua, Gordana Dermody, Leslie F. Sikos, Shamsul Islam

***Abstract*—**The advancement in information and communication technologies (ICT) and wireless sensor networks have played a pivotal role in the design and development of real-time healthcare solutions mainly targeting the elderly living in health-assistive smart homes. Such smart homes are equipped with sensor technologies to detect and record activities of daily living (ADL). This survey reviews and evaluates existing approaches and techniques based on real-time sensor-based modeling and reasoning in single-user and multi-user environments. It classifies the approaches into three main categories: learning-based, knowledge-based, and hybrid, and evaluates how they handle temporal relations, granularity, and uncertainty. The survey also highlights open challenges across various disciplines (including computer and information sciences, and health sciences) to encourage interdisciplinary research for the detection and recognition of ADLs and discusses future directions.

***Keywords*—**Daily Living Activities (ADL), smart homes, single-user environment, multi-user environment.

I. Introduction to Daily Living Activity Recognition in Smart Homes

THE advancement of information and communication Technologies (ICT) [1] and wireless sensor networks [2] led to the design and development of real-time information systems for decision making in several domains, such as e-commerce [3], e-education [4], manufacturing [5], security and surveillance [6], and many more. Healthcare is not an exception, and the development of ubiquitous environments for real-time patient monitoring has significantly increased patients' quality of life.

The World Health Organization (WHO) mentions that between 2015 and 2050, the sector of the population with more than 60 years will increase from 900 million to 2 billion [7]. Additionally, according to the Australian Law Reform Commission, the population aged 65 or over will be projected up to 22.6% by 2054–55[8]. As the number of the elderly expands, and considering the COVID-19 crisis, placing older adults in care centers may not be sustainable. However, older adults often suffer from chronic conditions that may impact their health and safety and require a certain level of clinical monitoring. Smart home sensors make this possible, and monitor cognitive changes such as dementia, and changes in health, which can be detected early by such a system. This sensing infrastructure features low-cost, low-power, and high-performance solutions. It has been estimated that smart homes can contribute about $7 trillion in cost reduction annually worldwide [9]. Moreover, this technology may support clinical decision-making of health professionals, i.e., doctors, nurses, or caregivers, and early recognition and triage in case of emergencies.

The real-time healthcare information system identifies and records the activities of daily living of patients in their homes. The term "activities of daily living" (ADL) was coined in the 1950s by Sidney Katz [10] who developed the Katz Index of Independence in Activities of Daily Living, also known as the Katz ADL. Furthermore, the researchers investigated and modeled different types of activities (atomic, complex, composite) with respect to time either as sequential, interleaved, and concurrent for single person environments, and parallel and collaborative activities for multi-user environments, as shown in Table I [11], [12].

TABLE I
ADL Recognition Methods

Activity Type	Definition	Example
Sequential	Each activity can be performed after another in a sequential manner	Making meals then washing dishes
Concurrent	More than one activity can be performed at a time by a user	Taking medicines while watching TV
Interleaved	Where a user can switch between activities	Switches between stirring soup and making a sandwich in the kitchen
Parallel	Many activities can be performed by multiple users at the same time	One user is cleaning the home while others are talking on the phone
Collaborative	Where multiple users perform activities together in a cooperative way	Making soup in the kitchen (one is chopping vegetables, the other is boiling water)

Significant activity recognition (AR) research has been conducted to identify ADLs. Some studies modeled contextual information using ontologies [13– 16] and machine learning approaches to perform probabilistic reasoning using Semantic Web technologies to detect simple and complex activities [17], [18] and anomalies in activities using neural networks [19]. Other studies found that events can be captured in data streams and that complex event processing and reasoning can be used to predict complex and composite activities [20]–[22]. Some considered the uncertainty of the data/conflicting preference to predict complex activities using answer set programming [23—24]. Generally, smart homes are limited to monitoring single users. However, more than one person may need to live in the home to provide assistance or companionship. In this regard, the literature shows that there is a great interest in a multi-user environment focusing on ADL activities where a composite activity is modeled. However, there are only a few researchers who track and identify multi-users in smart home environments [25], [26] by using machine learning algorithms, while many attempts to recognize activities individually, cooperatively, and in parallel by using Bayesian networks (BN), recurrent neural networks (RNN), conditional random fields (CRF), and Hidden Markov models (HMM) [27]–[30], and mobile applications [31].

This paper discusses various techniques for performing ADLs in the sensor-based single-and multi-user environment, their temporal nature, and how uncertainty is addressed. The main contributions of this paper can be outlined as follows:

(i) The current approaches are classified into three main categories: learning-based approaches, knowledge-based approaches, and hybrid approaches based on real-time sensor-based modeling and reasoning for ADLs in an activity recognition environment.

(ii) The existing methods and techniques are reviewed by classifying them based on whether they are single-user or multi-user to further evaluate how they handle temporal relations, granularity, and uncertainty.

Kulsoom Bughio is with the Edith Cowan University, Australia (e-mail: k.bughio@ecu.edu.au).

(iii) Based on the evaluation of different categories, open challenges are identified, and future directions are discussed.
The remainder of this paper is organized as follows. Section II discusses current approaches based on single-user and multi-user criteria. Section III provides a discussion about the findings of the literature review conducted. The key research challenges and future directions are described in Section IV. The last section V concludes this paper.

II. Current Approaches for ADL Recognition in Sensor-Based Environment

Many studies have been conducted for real-time sensor-based modeling and reasoning for ADLs. The current approaches can be classified into three main categories, as shown in Fig. 1

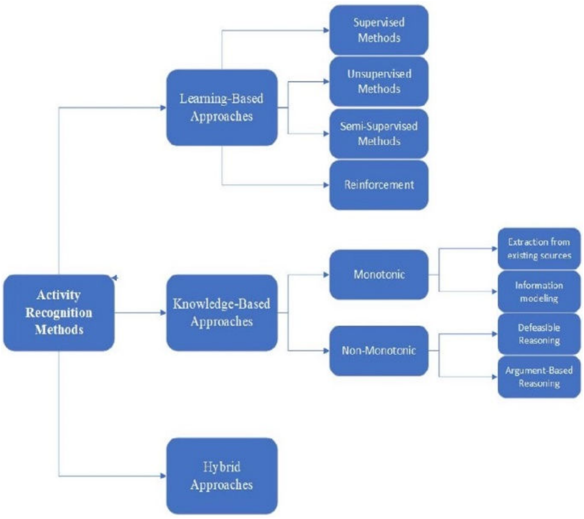


Fig. 1 ADL recognition methods

A. *Learning-Based Approaches*

The learning-based approaches, also known as data-driven approaches (DDA), are based on machine learning (ML) techniques. These methods develop models based on pre-existing datasets rather than intuition or individual experience. The benefit of learning-based methods is the ability to handle uncertainty. Some techniques, e.g., recurrent neural networks, also deal with temporal information [14]. In learning-based approaches, generally, models can be generated in four ways: supervised, unsupervised, semi-supervised, and reinforcement. In the following sections, these are discussed in detail.

1) Supervised ML Methods:

Most learning-based approaches employ supervised machine learning algorithms, which require a pre-existent labeled dataset of user behavior to infer the activity model. In these methods, the training data is used to train and test the model for accuracy and precision [16]. Once the model is trained, sensor data is used to detect physical activities such as gait speed, postures, physical motion, and health events. However, Mohmed [32] proposed another promising technology, namely, fuzzy finite state machine (FFSM), to handle uncertainty associated with human behavior. Bayesian networks, partially observable Markov decision processes (POMDP), Naïve Bayes, C4.5 decision trees, support vector machines (SVMs), Markov chains, hidden Markov models (HMMs), logistic regression, conditional random fields (CRFs), and neural networks (NNs) are some examples of supervised ML Methods.

By using supervised learning, most of the systems are dedicated to an activity recognition environment for single individuals [10], [33]; however, few researchers model activities in a multi-user environment. One of them is a benchmark study for multi-resident smart homes, which was presented by Tran et al. [27]. In this study, the researchers learned about the effectiveness of temporal learning algorithms and non-temporal learning by using sequential data and temporally manipulated features. ADL recognition was also performed [29]–[31], [34] in a multi-resident smart home based on Bayesian networks and a hidden Markov model. Chen et al. [28] proposed a system to recognize complex activities for the multi-resident environment by using both techniques such as HMM and CRF.

2) Unsupervised ML Methods:

Due to some limitations in the supervised approach, some studies investigate unsupervised approaches [35]–[37], which generate activity models relying on a training set of unlabeled sensor data. Unsupervised methods are divided into two types: clustering and association [38]. Most researchers used unsupervised clustering methods in a single-user environment to detect various physical activities, such as walking, sitting, standing, and jogging. For example, Lu et al., [36] provided a method for ADL classification by using a smartphone accelerometer. Negin et al. [35] proposed a framework for early diagnosis of cognitive impairments and ADL activity discovery and scene modeling in health care. While the location of multi-users in a home is detected by the vectorization approach using an algorithm named sMRT to track the location of each resident by using the clustering technique, thereby estimating the number of residents in the smart home with the help of event association [26].

3) Semi-Supervised Methods:

Semi-supervised learning methods improve the model computed through the training set by using unlabeled data [18]. In semi-supervised learning, both labeled and unlabeled data are used for training. There are some examples of semi-supervised methods, i.e., self-training, co-training, graph-based methods, expectation-maximization (EM) with generative mixture models, and transductive support vector machines [40].

4) Reinforcement:

Reinforcement learning (RL), also known as neurodynamic programming, is an approach for building agents automatically. These agents take a reward function after the performance measure. The system can direct its problem space and provide feedback in terms of rewards and punishments [41]. In activity recognition, Richard [42] proposed a recommendation system to guide patients diagnosed with Alzheimer's disease, in performing ADLs based on changing human mental state, behavior, and environmental contexts. Zhang et al. [43] proposed an effective way of recognizing human behavior activities in smart homes by using deep reinforcement learning (DL). Table II shows the various learning-based ML methods in single and multi-user environments for atomic and complex activities.

TABLE II
Summary of Learning-Based Approaches for ADL Detection and Recognition

Type of Learning	References	ADL Temporal Relation	Granularity Level of ADL	Uncertainty Handling
		Single-User		
Supervised	[44]	Concurrent	Complex	No
	[45]	Sequential, concurrent	Atomic, complex	No
	[32]	Sequential	Simple	Yes

Unsupervised	[46]	Sequential, interleaved, concurrent	Simple, complex	No
Reinforcement Learning	[42]	Sequential	Simple	Yes
		Multi-user		
Supervised	[28]	Sequential, parallel	Complex	No
	[31]	Parallel	Complex	No
	[34]	Parallel /cooperative	Complex	No
	[47]	Sequential, cooperative, collaborative	Complex	No
Unsupervised	[26]	Joint events	Atomic	No

B. Knowledge-Based Approaches

Knowledge-based approaches are also known as knowledge-driven approaches (KDA). In these approaches, "an activity model is developed through the incorporation of rich prior domain knowledge obtained from the application domain, using knowledge engineering and knowledge management techniques" [16]. These approaches use previous knowledge to create a logic-based recognition model, where activities can be modeled and expressed through numerous formats (i.e., rules, and ontologies) [14], [48] and can be mined from other sources (e.g., Web resources, or unlabeled datasets of activities) [18], that can be interpreted by humans and machines, rather than of the acquisition of labeled training datasets. The knowledge structure of KDAs can be modeled using various techniques, which are mainly classified as monotonic and non-monotonic. smart home with the help of event association [26].

1) Monotonic Logics:

Monotonic logic is also known as standard logic or classical logic, once something is true, it is true forever [49]. It is used as a specification language to represent declarative knowledge. However, it represents a monotonicity property based on the conclusion involved by a body of knowledge, if additional knowledge is added there is no effect on the conclusion, it remains valid.

a) Information extraction from existing sources

Different studies have been conducted using a mining-based approach [46], [50]. These approaches retrieve definitions of activities and phrases that define involved objects and the activity performance process by using information retrieval techniques. The system used in this approach is designated to overcome the deficiency of annotated ADL data problems and the variation or complex activity performance between various individuals [14]. The researchers [51] compute activity models by mining activity data from pictures and videos even sporadic activities are involved. An ambient assistant system named, ontology-based Ambient-Aware LIfeStyle tutoring for A BETter Health (ontoAALISABETH) [52] provides an ambient assisted living framework to monitor the lifestyle of older people, who are not suffering from major chronic diseases or severe disabilities. It integrates an ontology with a rule-based and a complex event processing (CEP) engine for supporting the timed reasoning but lack to deal with conflicting events.

b) Information modeling

In the logical-based approach, to present the knowledge about activity, different types of logical formalisms are used. In this regard, various methods were used to build a recognition model that infers actions or activity intentions through general knowledge-based rules, where training data was not required such as, a method to identify human intention based on percept sequence by using event calculus [53], tagged visual contents shared on the web [54]. In healthcare, the iKnow activity recognition system [13] for a single-user environment provides capabilities, which enhance the extraction of behavior patterns and behavior changes, including how activities are performed, idiosyncratic, and habitual knowledge. Triboan et al. [55] developed a multithreaded decision algorithm and a system prototype capturing generic knowledge and inhabitant-specific preferences for conducting composite ADLs to support the segmentation process. Another researcher, Bennasar et al. [14] proposed a system to support older people to live independently in their own homes by increasing the quality of life in terms of care, safety, and security. Ye et al. [56] proposed a novel approach for a multi-user environment to recognize concurrent activities, which are independent of sensor deployment and activities of interest.

2) Non-Monotonic Logics:

Non-monotonic logics have been described as "a kind of inference in which reasoners draw tentative conclusions, enabling reasoners to retract their conclusion(s) based on further evidence" [49]. In other words, facts and rules can be changed at any time, with some examples given below.

a) Defeasible Reasoning

This is a non-monotonic logic formalism that delivers instinctive knowledge representation and advanced resolution mechanisms such as inconsistent [57] and conflicting information (uncertainty) [58]. Defeasible reasoning performs better than classical logic in terms of computational complexity, dealing with incomplete information, and initiating new reasoning for non-accustomed users (doctors, patients, etc.). An implementation example is a ReDEF framework (Context-Aware Recognition of Interleaved Activities using OWL 2 and Defeasible Reasoning), which has been proposed [58] for ADL recognition and can be used for supporting the diagnosis of Alzheimer's disease in a controlled environment.

b) Argument-Based Reasoning

Argumentation theory, also known as common-sense reasoning, is when a person reasons about his/her activities by evaluating the potential results [49]. The argumentation-based approach is also used to deal with conflicting data, where preference arguments are built, defining which one will be a suitable solution, unlike a rule-based system, in which preferences are defined based on another for conflicting data. Guerrero et al. [57] used two semantics for ELP Answer set semantics and well-founded semantics. They aimed to follow answer set programming roots for capturing negation as failure in an argumentation-based setting. Based on human motives, goals, and prioritized actions, Guerrero et al. [59] proposed an argument-based approach for tracking and monitoring old people's complex activities. Table III shows the ADL recognition performed by different researchers by using knowledge-based techniques such as monotonic and non-monotonic techniques for various types of activities (simple, complex, and composite).

C. Hybrid Approaches

By combining DDA and KDA (hybrid approaches), some limitations of previous approaches can be overcome. In DDA, the activities can be represented by probabilistic and statistical models [13]. This is useful for dealing with noise and uncertainty of sensor measurements, however, semantic relationships between sensor events and activities cannot be captured. In contrast, KDA approaches capture complex semantic relationships, but these approaches are often too rigid to cope with noise and uncertainty [60].

TABLE III
SUMMARY OF EXISTING KNOWLEDGE-BASED APPROACHES

Type of Learning	References	ADL Temporal Relation	Granularity Level of ADL	Uncertainty Handling
		Single-User		
Monotonic	[14]	Sequential	Complex	No
	[55]	Parallel, concurrent, incremental	Simple, composite	No
	[13]	Interleaved	Situation descriptor	No
	[53]	Percept sequence	Complex intensions	Yes
	[52]	Sequential	Complex	No
Non-Monotonic	[58]	Sequential, interleaved, experimental	Complex	Yes
	[49]	Activity-Goal	Complex	Yes
		Multi-User		
Monotonic	[56]	Concurrent	Complex	No

In the context of a single-user environment, hybrid approaches have been developed that can fully utilize unsupervised techniques combined with ontologies for activity recognition. The initial model is grown by combining KDA's expressivity and DDA's uncertainty handling; in this regard, some researchers [17], [18], [37], [61] used a combination of ontology and probabilistic reasoning, i.e., Markov logic network, to recognize sequential, concurrent, and interleaved activities.

To recognize complex ADLs in a multi-resident environment, a hybrid approach was modeled by Roy et al. [31]. In this model, two extremes of sensors, i.e., ambient sensors and wearable sensors were used to capture the users' spatiotemporal behavior and location. Nguyen et al. [62] provided a mechanism for activity segmentation and recognition in the context of multi-resident homes, using ontology and unsupervised machine learning strategies, such as pattern discovery. Table IV explains the different hybrid techniques, i.e., probabilistic ontology modeled for activity recognition in a sequential, interleaved, and concurrent manner for complex and composite activities.

TABLE IV
SUMMARY OF EXISTING HYBRID APPROACHES

Technique	References	ADL Temporal Relation	Granularity Level of ADL	Uncertainty Handling
		Single-User		
Ontology +Temporal Formalism	[63]	Sequential, concurrent	Composite	NO
Ontology + MLN	[18]	Sequential, interleaved	Complex	Yes
Probabilistic Ontology	[17]	Sequential, interleaved, concurrent	Composite	Yes
Probabilistic Ontology	[61]	Sequential	Complex	Yes
		Multi-User		
Ontology + Unsupervised ML	[62]	Concurrent	Complex	No

III. DISCUSSION

Significant progress has been made for modeling, reasoning, and prediction of Activities of Daily Living (ADLs) of persons in their homes [13], [14]. Among the approaches investigated, a few, such as MLNs, ontologies, argumentation, and CEP engine, are promising that attempt to recognize ADLs in a single user environment with simple or complex activities. This survey paper discusses existing techniques for modeling and reasoning ADLs in an activity recognition environment and provides a taxonomy for their classification that includes learning-based methods, knowledge-driven approaches, and hybrid approaches. Table II shows that most of the learning-based methods are used to generate models for modeling atomic/simple and complex ADL activities in both single and multi-user environments. Some of these approaches [32], [42] used for ADL modeling in a single-user environment can handle uncertainty using fuzzy sets and learning automata, however, they are limited to modeling only sequential activities. Modeling approaches for ADLs in a multi-user environment can model complex activities, however, first, they fail to model composite activities, and second, they lack the capability of handling uncertainty for ADL activities modeling and predication when the underlying information is incomplete and/or contradictory. Although limited systems can identify health events for ADLs, the modeling is limited to a single user.

Table III shows that most of the monotonic logic-driven knowledge-based systems can model simple and complex model ADL activities in a single-user environment except the [55] approach, which can model composite activities. However, they can handle incomplete information using weights [53], but not contradictory as they are not able to deal with uncertainty. Additionally, a non-monotonic logic-based system for the single-user environment can model and reason about sequential and complex activities and perform support modeling and reasoning with incomplete and contradictory knowledge by using defeasible logic and answer set programming. However, these systems are unable to model composite activities and they use a single criterion to resolve contradictory knowledge. In a multi-user environment, [56] complex activities can be modeled by applying monotonic logic, but they are not able to model composite activities and cannot handle uncertainty.

Significant work to model activities through hybrid approaches using ontologies and ML techniques is shown in Table IV. These approaches can model and reason concurrent and interleaved activities in both single-user and multi-user environments. while the only single-user approaches model and reason with uncertainty. Ontologies perform the best when it comes to expressiveness and probabilistic knowledge is used in these systems. Researchers have used Markov models, fuzzy sets, Dempster Shafer theory to deal with uncertainty in a single-user environment for sequential, interleaved, and concurrent activities.

IV. RESEARCH CHALLENGES AND FUTURE RESEARCH DIRECTIONS

The existing techniques and systems have significant limitations in recognizing composite activities for multi-users in real-time systems, especially when data is coming from various sources and if the data is incomplete and/or contradictory. First, the existing approaches can model only simple or complex activities and they fail to consider multiple and composite activities [55] in real-time environments, such as a person with dementia living with his partner and doing activities in parallel and collaboratively. Second, the home context can be dynamic if more than one person lives in the home, or if the person has frequent visitors. If there are pets in the house, they can further increase complexity. In environments this dynamic, it can be challenging to detect and recognize the sensor-captured ADLs due to their unstructured and complex nature along with the number of individuals, infinite variations in speed, and style of performance [35]. Third, the sensor data is derived from different sources in a particular time frame that is uncertain. Measuring the uncertainty of events using knowledge-based reasoning (via ontologies) often does not produce accurate results [17]. Moreover, due to age-related changes in health and the worsening of chronic diseases, the quality of life of the elderly may fluctuate and decline over time, with temporary periods of stabilization. Monitoring ADLs together with any signs of deteriorating health and functional decline is important to be able to provide supportive care early on to prevent adverse health outcomes.

To overcome these limitations, there is a need to design and develop rules and ontologies to model composite activities in a real-time system for the ADLs in multi-user environments. In modeling and reasoning ADLs, the system would be able to understand what happens in terms of contradictory events coming from various sources, integrate them, and reason about events. To handle incomplete/or contradictory information in real-time coming from various sources, an argumentation approach could be used to capture and reason with ADL activities with the CEP engine such as ETALIS, which processes multiple streams of data and reasons about them. This could also provide a real-time monitoring system to alert patients, caregivers, and doctors about cognitive behavior changes for patients.

V. CONCLUSION

Activity is a pivotal task for the identification of ADLs in smart home environments. To detect and monitor ADLs and assist individuals to secure their safety and well-being, different types of activities (atomic, complex, and composite) are investigated by researchers. This can provide advanced monitoring for smart home environments by utilizing technologies for which clinicians have worked with engineers and computer scientists to obtain clinically relevant ground truth to model potential health events. This paper is a systematic review of ADLs in smart home environments, highlighting challenges in terms of composite activities in a multi-user environment and conflicting data in real-time. This survey identified limitations of the state of the art and suggested directions for future research.

REFERENCES

[1] R. Mansell, “Information and Communication Technology Policy Research in the United Kingdom: A Perspective,” Canadian Journal of Communication, vol. 19, no. 1, pp. 1–11, 1994, doi: 10.22230/cjc.1994v19n1a792.

[2] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, “Wireless sensor networks: A survey,” Computer Networks, vol. 38, pp. 393–422, 2002, doi: 10.1109/WAINA.2009.192.

[3] W. H. DeLone and E. R. McLean, “Measuring e-commerce success: Applying the DeLone and McLean Information Systems Success Model,” International Journal of Electronic Commerce, vol. 9, no. 1, pp. 31–47, 2004, doi: 10.1080/10864415.2004.11044317.

[4] A. Shimada, S. Konomi, and H. Ogata, “Real-time learning analytics system for improvement of on-site lectures,” Interactive Technology and Smart Education, vol. 15, no. 4, pp. 314–331, 2018, doi: 10.1108/ITSE-05-2018-0026.

[5] F. Nawaz, N. K. Janjua, and O. K. Hussain, “PERCEPTUS: Predictive complex event processing and reasoning for IoT-enabled supply chain,” Knowledge-Based Systems, vol. 180, no. May, pp. 133–146, 2019, doi: 10.1016/j.knosys.2019.05.024.

[6] S. Srinivasan, H. Latchman, J. Shea, T. Wong, and J. McNair, “Airborne traffic surveillance systems - Video surveillance of highway traffic,” Proceedings of the ACM Second International Workshop on Video Sureveillance and Sensor Networks, pp. 131–135, 2004.

[7] G. Jorge and F. M. Mena, “OPAIEH : An Ontology-based Platform for Activity Identification of the Elderly at Home,” vol. 24, no. 2, pp. 481–495, 2020, doi: 10.13053/CyS-24-2-3373.

[8] Australian Government; Australian Law Reform Commission, “Who are older Australians?,” pp. 1–8, 2017, [Online]. Available: https://www.alrc.gov.au/publication/elder-abuse-a-national-legal-response-alrc-report-131/2-concepts-and-context-2/who-are-older-australians/.

[9] N. Oukrich, “Daily Human Activity Recognition in Smart Home based on Feature Selection, Neural Network and Load Signature of Appliances,” 2019, [Online]. Available: https://hal.archives-ouvertes.fr/tel-02193228.

[10] Peter F.; Deb L.; Sukesh S.; Shoshana B., “Activities of Daily Living (ADLs),” Encyclopedia of Disability, pp. 1–6, doi: 10.4135/9781412950510.n12.

[11] A. Benmansour, A. Bouchachia, and M. Feham, “Multioccupant activity recognition in pervasive smart home environments,” ACM Computing Surveys, vol. 48, no. 3, 2015, doi: 10.1145/2835372.

[12] Q. Li, R. Gravina, Y. Li, S. H. Alsamhi, F. Sun, and G. Fortino, “Multi-user activity recognition: Challenges and opportunities,” Information Fusion, vol. 63, no. March, pp. 121–135, 2020, doi: 10.1016/j.inffus.2020.06.004.

[13] G. Meditskos and I. Kompatsiaris, “iKnow: Ontology-driven situational awareness for the recognition of activities of daily living,” Pervasive and Mobile Computing, vol. 40, pp. 17–41, 2017, doi: 10.1016/j.pmcj.2017.05.003.

[14] M. Bennasar et al., “Knowledge-Based Architecture for Recognising Activities of Older People,” 23rd International Conference on Knowledge-Based and Intelligent Information & Engineering Systems, vol. 159, pp. 590–599, 2019, doi: 10.1016/j.procs.2019.09.214.

[15] A. G. Salguero, M. Espinilla, P. Delatorre, and J. Medina, “Using ontologies for the online recognition of activities of daily living,” Sensors (Switzerland), vol. 18, no. 4, pp. 1–22, 2018, doi: 10.3390/s18041202.

[16] A. G. Salguero, J. Medina, P. Delatorre, and M. Espinilla, “Methodology for improving classification accuracy using ontologies: application in the recognition of activities of daily living,” Journal of Ambient Intelligence and Humanized Computing, vol. 10, no. 6, pp. 2125–2142, 2019, doi: 10.1007/s12652-018-0769-4.

[17] K. S. Gayathri, K. S. Easwarakumar, and S. Elias, “Probabilistic ontology based activity recognition in smart homes using Markov Logic Network,” Knowledge-Based Systems, vol. 121, pp. 173–184, 2017, doi: 10.1016/j.knosys.2017.01.025.

[18] G. Civitarese, C. Bettini, T. Sztyler, D. Riboni, and H. Stuckenschmidt, “newNECTAR: Collaborative active learning for knowledge-based probabilistic activity recognition,” Pervasive and Mobile Computing, vol. 56, pp. 88–105, 2019, doi: 10.1016/j.pmcj.2019.04.006.

[19] D. Arifoglu and A. Bouchachia, “Detection of abnormal behaviour for dementia sufferers using Convolutional Neural Networks,” Artificial Intelligence in Medicine, vol. 94, no. May 2018, pp. 88–95, 2019, doi: 10.1016/j.artmed.2019.01.005.

[20] Y. Liu, X. Wang, Z. Zhai, R. Chen, B. Zhang, and Y. Jiang, “Timely daily activity recognition from headmost sensor events,” ISA Transactions, vol. 94, pp. 379–390, 2019, doi: 10.1016/j.isatra.2019.04.026.

[21] E. Wu, Y. Diao, and S. Rizvi, “High-performance complex event processing over streams,” Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 407–418, 2006, doi: 10.1145/1142473.1142520.

[22] K. Taylor and L. Leidinger, “Ontology-Driven complex event processing in heterogeneous sensor networks,” Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 6643 LNCS, no. PART 2, pp. 285–299, 2011, doi: 10.1007/978-3-642-21064-8_20.

[23] E. Guerrero, J. C. Nieves, and H. Lindgren, “Semantic-based construction of arguments: An answer set programming approach,”

International Journal of Approximate Reasoning, vol. 64, pp. 54–74, 2015, doi: 10.1016/j.ijar.2015.06.009.

[24] J. C. Nieves, S. Partonia, E. Guerrero, and H. Lindgren, “A probabilistic non-monotonic activity qualifier,” Procedia Computer Science, vol. 52, no. 1, pp. 420–427, 2015, doi: 10.1016/j.procs.2015.05.007.

[25] P. Lapointe, “A New Device to Track and Identify people in a Multi-Residents Context,” vol. 00, 2020, doi: 10.1016/j.procs.2020.03.082.

[26] T. Wang and D. J. Cook, “sMRT: Multi-Resident Tracking in Smart Homes with Sensor Vectorization,” IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1–1, 2020, doi: 10.1109/tpami.2020.2973571.

[27] S. N. Tran et al., “On multi-resident activity recognition in ambient smart-homes,” Artificial Intelligence Review, 2019, doi: 10.1007/s10462-019-09783-8.

[28] R. Chen and Y. Tong, “A Two-stage Method for Solving Multi-resident Activity Recognition in Smart Environments,” pp. 2184–2203, 2014, doi: 10.3390/e16042184.

[29] L. Fu and J. Y. Hsu, “Interaction models for multiple-resident activity recognition in a smart home,” pp. 3753–3758, 2010.

[30] G. Singla and D. J. Cook, “Recognizing independent and joint activities among multiple residents in smart environments,” pp. 57–63, 2010, doi: 10.1007/s12652-009-0007-1.

[31] N. Roy, A. Misra, D. Cook, and D. Cook, “Ambient and smartphone sensor assisted ADL recognition in multi-inhabitant smart environments,” Journal of Ambient Intelligence and Humanized Computing, vol. 7, no. 1, pp. 1–19, 2016, doi: 10.1007/s12652-015-0294-7.

[32] G. Mohmed, A. Lotfi, and A. Pourabdollah, “Enhanced fuzzy finite state machine for human activity modelling and recognition,” Journal of Ambient Intelligence and Humanized Computing, no. 0123456789, 2020, doi: 10.1007/s12652-020-01917-z.

[33] P. Asghari, E. Soleimani, and E. Nazerfard, “Online human activity recognition employing hierarchical hidden Markov models,” Journal of Ambient Intelligence and Humanized Computing, vol. 11, no. 3, pp. 1141–1152, 2020, doi: 10.1007/s12652-019-01380-5.

[34] A. Benmansour, A. Bouchachia, and M. Feham, “Modeling interaction in multi-resident activities,” Neurocomputing, vol. 230, no. May 2016, pp. 133–142, 2017, doi: 10.1016/j.neucom.2016.05.110.

[35] F. Negin and F. Brémond, “An unsupervised framework for online spatiotemporal detection of activities of daily living by hierarchical activity models,” Sensors (Switzerland), vol. 19, no. 19, pp. 1–34, 2019, doi: 10.3390/s19194237.

[36] Y. Lu, Y. Wei, L. Liu, J. Zhong, L. Sun, and Y. Liu, “Towards unsupervised physical activity recognition using smartphone accelerometers,” Multimedia Tools and Applications, vol. 76, no. 8, pp. 10701–10719, 2017, doi: 10.1007/s11042-015-3188-y.

[37] D. Riboni, T. Sztyler, G. Civitarese, and H. Stuckenschmidt, “Unsupervised recognition of interleaved activities of daily living through ontological and probabilistic reasoning,” UbiComp 2016 - Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 1–12, 2016, doi: 10.1145/2971648.2971691.

[38] P. B. Amelia, “Unsupervised Machine Learning: What is , Algorithms , Example What is Unsupervised Learning ?”

[39] U. Machine and L. Algorithms, “Supervised and Unsupervised Machine Learning Algorithms Get your FREE Algorithms Map,” pp. 1–45, 2020.

[40] X. Zhu, “Semi-Supervised learning literature survey.” 2005.

[41] S. Mishra, “Unsupervised Learning and Data Clustering,” Towards Data Science, pp. 1–16, 2017.

[42] R. O. Oyeleke, C. Y. Yu, and C. K. Chang, “Situ-Centric Reinforcement Learning for Recommendation of Tasks in Activities of Daily Living in Smart Homes,” Proceedings - International Computer Software and Applications Conference, vol. 2, pp. 317–322, 2018, doi: 10.1109/COMPSAC.2018.10250.

[43] W. W. Zhang and W. Li, “A deep reinforcement learning based human behavior prediction approach in smart home environments,” Proceedings - 2019 International Conference on Robots and Intelligent System, ICRIS 2019, pp. 59–62, 2019, doi: 10.1109/ICRIS.2019.00024.

[44] T. Y. Wu, C. C. Lian, and J. Y. J. Hsu, “Joint recognition of multiple concurrent activities using factorial conditional random fields,” AAAI Workshop - Technical Report, vol. WS-07-09, pp. 82–87, 2007.

[45] L. Liu, Y. Peng, S. Wang, M. Liu, and Z. Huang, “Complex activity recognition using time series pattern dictionary learned from ubiquitous sensors,” Information Sciences, vol. 340–341, pp. 41–57, 2016, doi: 10.1016/j.ins.2016.01.020.

[46] T. Gu, Z. Wu, X. Tao, H. K. Pung, and J. Lu, “epSICAR: An emerging patterns-based approach to sequential, interleaved and concurrent activity recognition,” 7th Annual IEEE International Conference on Pervasive Computing and Communications, PerCom 2009, 2009, doi: 10.1109/PERCOM.2009.4912776.

[47] Y. T. Chiang, K. C. Hsu, C. H. Lu, L. C. Fu, and J. Y. J. Hsu, “Interaction models for multiple-resident activity recognition in a smart home,” IEEE/RSJ 2010 International Conference on Intelligent Robots and Systems, IROS 2010 - Conference Proceedings, pp. 3753–3758, 2010, doi: 10.1109/IROS.2010.5650340.

[48] G. Civitarese, T. Sztyler, D. Riboni, C. Bettini, and H. Stuckenschmidt, “POLARIS: Probabilistic and Ontological Activity Recognition in Smart-homes,” IEEE Transactions on Knowledge and Data Engineering, vol. PP, no. c, pp. 1–1, 2019, doi: 10.1109/tkde.2019.2930050.

[49] E. G. Rosero and P. D. Thesis, Representing and Reasoning about Complex Human Activities - an Activity-Centric Argumentation-Based Approach. Printed by Print & Media, Ume˚a University, 2016., 2016.

[50] F. Ongenae et al., “A probabilistic ontology-based platform for self-learning context-aware healthcare applications,” Expert Systems with Applications, vol. 40, no. 18, pp. 7629–7646, 2013, doi: 10.1016/j.eswa.2013.07.038.

[51] D. Riboni and M. Murtas, “Sensor-based activity recognition: One picture is worth a thousand words,” Future Generation Computer Systems, vol. 101, pp. 709–722, 2019, doi: 10.1016/j.future.2019.07.020.

[52] R. Culmone, P. Giuliodori, and M. Quadrini, “Human Activity Recognition using a Semantic Ontology-Based Framework,” International Journal on Advances in Intelligent Systems, vol. 8, no. 2, pp. 159–168, 2015, [Online]. Available: http://www.iariajournals.org/intelligent_systems/www.iaria.org.

[53] J. Kim, M. Jeon, H. Park, S. Bae, S. Bang, and Y. Park, “An approach for recognition of human ’ s daily living patterns using intention ontology and event calculus,” vol. 132, pp. 256–270, 2019, doi: 10.1016/j.eswa.2019.04.004.

[54] D. Riboni and M. Murtas, “Sensor-based activity recognition: One picture is worth a thousand words,” Future Generation Computer Systems, vol. 101, pp. 709–722, 2019, doi: 10.1016/j.future.2019.07.020.

[55] D. Triboan, L. Chen, F. Chen, and Z. Wang, “A semantics-based approach to sensor data segmentation in real-time Activity Recognition,” Future Generation Computer Systems, vol. 93, pp. 224–236, 2019, doi: 10.1016/j.future.2018.09.055.

[56] J. Ye and G. Stevenson, “Semantics-driven multi-user concurrent activity recognition,” Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 8309 LNCS, no. 256873, pp. 204–219, 2013, doi: 10.1007/978-3-319-03647-2-15.

[57] E. Guerrero, J. C. Nieves, and H. Lindgren, “Semantic-based construction of arguments: An answer set programming approach,” International Journal of Approximate Reasoning, vol. 64, pp. 54–74, 2015, doi: 10.1016/j.ijar.2015.06.009.

[58] G. Meditskos, E. Kontopoulos, and I. Kompatsiaris, “ReDef: Context-aware recognition of interleaved activities using OWL 2 and defeasible reasoning,” CEUR Workshop Proceedings, vol. 1488, pp. 31–42, 2015.

[59] E. Guerrero, J. C. Nieves, and H. Lindgren, “An activity-centric argumentation framework for assistive technology aimed at improving health,” Argument and Computation, vol. 7, no. 1, pp. 5–33, 2016, doi: 10.3233/AAC-160004.

[60] G. Civitarese, “Human Activity Recognition in Smart-Home Environments for Health-Care Applications,” pp. 1–1, 2019, doi: 10.1109/percomw.2019.8730719.

[61] M. H. M. Noor, Z. Salcic, and K. I. K. Wang, “Enhancing ontological reasoning with uncertainty handling for activity recognition,” Knowledge-Based Systems, vol. 114, no. May 2018, pp. 47–60, 2016, doi: 10.1016/j.knosys.2016.09.028.

[62] D. Nguyen, L. Nguyen, and S. Nguyen, “A Novel Approach of Ontology-Based Activity Segmentation and Recognition Using Pattern Discovery in Multi-resident Homes,” Advances in Intelligent Systems and Computing, vol. 1013, pp. 167–178, 2020, doi: 10.1007/978-981-32-9186-7_19.

[63] G. Okeyo, L. Chen, and H. Wang, “Combining ontological and temporal formalisms for composite activity modelling and recognition in smart homes,” Future Generation Computer Systems, vol. 39, pp. 29–43, 2014, doi: 10.1016/j.future.2014.02.014.

Kulsoom S. BUGHIO was born in Thatta city in Sindh province of Pakistan. Early education earned at Thatta, after that earned BS (Honors) and M.Phil. (Master of Philosophy) degree in Computer Science from Institute of Mathematics and Computer Science, University of Sindh, Pakistan, currently doing Doctor of Philosophy (PhD) in Computing and Security, School of Science, Edith Cowan University, WA, Australia.

She has more than ten years of experience in academia and research. She taught various subjects of Computer science. She is a good researcher in the field of Computing where she has published publications during MPhil study independently and collaboratively. She has also been awarded an HEC-ECU joint scholarship in 2019 for pursuing PhD from Australia for four years. Along with this, she has participated in different pieces of training, internships, and certifications.

Ms BUGHIO is a member of GPN (Girls Programming Network) Western Australia since 2020. Ms BUGHIO is also a Postgraduate Representative for Research students for School of Science, Edith Cowan University this year and doing a casual academia job in the same School.

Science, Technology, Engineering, Mathematics Based Entrepreneurship Training within a Learning Company

Assoc. Prof. D. Mitova, PhD Diana Mitova, Doctoral student K. Mitrev

***Abstract*—** To prepare the current generation for the future, education systems need to change. It implies a new way of learning that is relevant to the demands of the times and the environment in which we live. The traditional approach of teaching individual subjects in isolation no longer meets the challenges of today's world, society, and work environment. Students need not only theoretical knowledge but transferable skills that will help them to become inventors and entrepreneurs, implement new ideas and create innovations. Science, technology, engineering, and mathematics education, better known as STEM, is now a real necessity for modern schools.

The subject of this study and entrepreneurship education, within a learning company with the application of STEM - technology, which encourages students to think outside the traditional box. STEM learning focuses the teacher's efforts on creating a model of entrepreneurial thinking and behavior in students and helping them solve problems in the world of business and entrepreneurship. Special attention is given to team activities and group work, simulation, experimentation, discussion and discourse, brainstorming, case studies, and project work.

Learning based on the implementation of various STEM projects in after-school activities, experiential learning, and an interdisciplinary approach are the means through which the educators better connect the local community and private businesses. Students learn to be a creative, experiment and take risks, and work in teams- the leading characteristics of every innovator and future entrepreneur. This article presents some European policies on STEM and entrepreneurship education. It also shares best practices for learning company training with the integration of STEM in the learning company training environment.

***Keywords*—**STEM, entrepreneurship, learning company, extracurricular activities.

Diana Mitova is with the South-West University “Neofit Rilski”, Blagoevgrad, Republic of Bulgaria (e-mail: didimitova2006@abv.bg)

Krassimir Mitrev is with the South-West University “Neofit Rilski”, Blagoevgrad, Republic of Bulgaria (e-mail: kr.mitrevzpg@gmail.com)

STEM (Science – Technology – Engineering – Mathematics) -based entrepreneurship training, within a learning company

Assoc. Prof. D. Mitova, Doctoral student K. Mitrev

***Abstract*—** To prepare the current generation for the future, education systems need to change. It implies a way of learning that meets the demands of the times and the environment in which we live. Productive interaction in the educational process implies an interactive learning environment and the possibility of personal development of learners based on communication and mutual dialogue, cooperation and good partnership in decision-making. Students need not only theoretical knowledge, but transferable skills that will help them to become inventors and entrepreneurs, to implement ideas. STEM education , is now a real necessity for the modern school. Through learning in a "learning company", students master examples from classroom practice, simulate real life situations, group activities and apply basic interactive learning strategies and techniques. The learning company is the subject of this study, reduced to entrepreneurship training in STEM - technologies that encourage students to think outside the traditional box. STEM learning focuses the teacher's efforts on modeling entrepreneurial thinking and behavior in students and helping them solve problems in the world of business and entrepreneurship. Learning based on the implementation of various STEM projects in extracurricular activities, experiential learning, and an interdisciplinary approach are means by which educators better connect the local community and private businesses. Learners learn to be creative, experiment and take risks and work in teams - the leading characteristics of any innovator and future entrepreneur. This article presents some European policies on STEM and entrepreneurship education. It also shares best practices for training company training , with the integration of STEM in the learning company training environment. The main results boil down to identifying some advantages and problems in STEM entrepreneurship education. The benefits of using integrative approaches to teach STEM within a training company are identified, as well as the positive effects of project-based learning in a training company using STEM. Best practices for teaching entrepreneurship through extracurricular activities using STEM within a training company are shared. The following research methods are applied in this research paper:

- Theoretical and comparative analysis of principles and policies of European Union countries and Bulgaria in the field of entrepreneurship education through a training company. Experiences in entrepreneurship education through extracurricular activities with STEM application within a training company are shared.

- A questionnaire survey to investigate the motivation of secondary vocational school students to learn entrepreneurship through a training company and their readiness to start their own business after completing their education.

Within the framework of learning through a "learning company" with the integration of STEM, the activity of the teacher-facilitator includes the methods: counseling, supervising and advising students during work. The expectation is that students acquire the key competence "initiative and entrepreneurship" and that the cooperation between the vocational education system and the business in Bulgaria is more effective.

***Keywords*—**STEM, entrepreneurship, training company, extracurricular activities.

I. Introduction

THE application of the STEM (Science, Technology, Engineering, Math) concept in education aims to develop the skills of successful 21st century citizens. The skillful blending of science and technology in a learning environment builds highly adaptable and mobile young people who are able to communicate and work in teams, display creative curiosity, initiative and a critical eye for the problems of the day. STEM integrates scientific approaches that help students quickly and easily master new real-life knowledge, skills, and competencies, inspiring them to think outside the box, experiment and be creative.

The use of STEM - educational technologies, from an early age, stimulates the development of initial life skills, provokes creative curiosity and observation, problem-solving skills from the immediate environment. By applying the STEM approach, children learn to be active, to observe, explore, explain, not only to memorize, but also to creatively reproduce and use the information obtained. As adolescents grow older, the application of STEM in an educational setting directs their interest towards the sciences related to the professions of the future, towards the realization of independent scientific research, experimentation, and innovation. The entrepreneurial activity of students is stimulated in the direction of the emergence of business ideas and their effective implementation in practice.

II. European STEM policies and entrepreneurship education

Achieving high-quality education for all young people is a major concern for Europe's future. In 2020, the European Commission adopted a number of documents focusing on the creation of the European Education Area, a European Skills Agenda, a renewed VET policy and a European Research Area [16].

Diana Mitova is with the South-West University “Neofit Rilski”, Blagoevgrad, Republic of Bulgaria (e-mail: **didimitova2006@abv.bg)**

Krassimir Mitrev is with the South-West University “Neofit Rilski”, Blagoevgrad, Republic of Bulgaria (e-mail: **kr.mitrevzpg@gmail.com** kr.mitrev@swu.bg **)**

It outlines how cooperation can improve the quality, inclusion, and digital and environmental dimensions of Member States' education systems. It sets out the means and milestones for achieving the European Education Area by 2025, with the support of the plan for the renewal of Europe (Next Generation EU) and the Erasmus+ program. Additional investment in education will contribute to the expansion of innovative learning models and provide a modern learning environment for STEM applications.

Learning through the methods of science, technology, engineering, and mathematics, includes interactive environments, the establishment of centers for the personal development of students and young people, the creation and development of centers of excellence in vocational education and training on a territorial or sectoral basis, the introduction and use of scientific approaches, innovation and artificial intelligence systems in education and training [14].

EU STEM is a European network of national STEM platforms, bringing together the national STEM strategies of each of the European Union Member States. The European Partner Organizations have been set up with the aim of close cooperation between government, education, and industry of individual countries. EU STEM ensures the development of new STEM strategies and the exchange of good practices between national STEM platforms. EU STEM implements close cooperation with a multitude of European, national, and regional partners, including national and regional governments, industry and institutions at European level [12].

ARTIFEX is an Erasmus+ project funded by the European Commission to create a practical European framework for learning in science, technology, engineering, and mathematics (STEM) in innovative learning environments outside the classroom, such as FabLabs or Makerspaces that produce products using modern technology [2].

The Do Well Science Coalition was established in line with the key priorities set by the European Commission [10].

In summary, the priorities boil down to the following guidelines:

- improving academic knowledge and acquired skills for lifelong learning, including for disadvantaged students;
- applying an innovative and integrated approach to teaching;
- applying an approach based on the digitization of content;
- improving the quality of education by using and adapting modern active learning methods;
- student-centered learning with the application of new teaching and learning technologies;
- overcoming disparities and early school leaving;
- the use of an interdisciplinary approach in combining knowledge from science, technology and the natural sciences, which stimulates students' innovative and critical thinking;
- sharing good practices, exploring, and adapting new STEM strategies and educational technologies [13].

Bulgaria has been part of the European STEM Coalition since 2017, implementing its national STEM strategy. The need for innovation in pedagogical interaction and the teaching-learning process is the main message in our European and national education policy. Emphasis is placed on the ability to discover and solve problems, to compare, analyze, systematize, and summarize scientific facts and information, and to use information and communication technologies (ICT) to unleash the technical and creative potential of learners. The desired changes in the quality of school education in science and technology can be achieved by implementing innovative teaching practices. The use of new technologies with the integration of science in the classroom, the application of integrated science teaching and learning practices in out-of-school settings, enhances teachers' motivation for learning and professional development.

III. STEM LEARNING IN ENTREPRENEURIAL TRAINING – ADVANTAGES AND PROBLEMS

"Initiative and entrepreneurship" is one of the key competences for lifelong learning, and entrepreneurship education is an important policy objective of member states. For adolescents to become innovators in their chosen career field, it is important to develop their entrepreneurial skills. Entrepreneurship training shapes a new way of thinking in young people, providing them with the knowledge, skills and attitudes for entrepreneurial activity and helping to develop their entrepreneurial culture.

School education is increasingly adopting a cross-curricular approach to acquire basic skills in STEM - science, technology, engineering, and mathematics education [4].

STEM is based on constructivism as an educational philosophy. It is student-centered and focuses on the student and his or her individual educational needs. Learners are at the center of the educational process and play the role of active subjects, independently mastering practical experiences. The learning process is accompanied by idea generation, solution seeking and problem solving, while participating in real life situations. The teacher advises, monitors, motivates, and provides support when needed.

The main advantages of STEM education, interpreted in the context of entrepreneurial training, boil down to the following aspects:

- bringing together knowledge from different subjects in an integrated curriculum based on cross-curricular links;
- learners gain an objective view of nature, society and the world in which they live, increasing their flexibility, adaptability and creativity in solving the problems of the day;
- understand and make sense of scientific concepts, facts and patterns among the processes and phenomena studied, compare quantities, explore relationships, and defend different points of view in collaborative solutions;
- apply new knowledge to concrete practical situations, master generalized means of activity and develop life skills;
- increase their motivation, experience pleasure and satisfaction from their activity, gain practical experience;

- acquire a wide range of social competences and skills in the field of modern technology and entrepreneurial activity;
- discover new life perspectives and career development opportunities.

STEM education builds a logical link between individual subjects and entrepreneurial learning, through students' participation in independent research, and group projects with an entrepreneurial focus. Through the skillful combination of experiments and games, students acquire the skills to investigate, analyze, summarize, and compare facts and information, drawing their own conclusions and inferences.

It is an important requirement that schools teaching STEM entrepreneurship have specialized facilities and that teachers are experts in their field. To this end, specialized school centers with a STEM focus are being established. They provide the necessary conditions and interactive learning environments for quality STEM education in schools [5].

The National Program "Building School STEM Environments" finances projects for building school STEM centers in Bulgaria. It aims to increase students' interest and achievement in science and technology.

Some authors (Mirtschewa, 2021) analyze some problems that may accompany STEM education. These are most often associated with a misunderstanding of the teacher's role in the learning process. Such are the cases where more emphasis is placed on teacher-centered teaching, despite the requirement in STEM learning environments that the teacher should primarily be a mentor and facilitator. There is also the opposing view that children do not need a teacher, but should explore the world for themselves, and that the introduction of new technologies into education can replace the teacher. Another problem stems from the perception of the more abstract nature of knowledge and that STEM learning is most effective through rote memorization and reproduction of more complex scientific facts and concepts. This runs counter to the core STEM concept where children learn to think and act like scientists, mathematicians, engineers, and inventors by gaining their own experiences. Another common problem is that the end product of the activity is primarily evaluated without sufficient attention to the process of its creation. There is more emphasis on the fun side of STEM and less on the cognitive side. Another commonly reported problem is the insufficient time allocated to STEM activities, which does not allow for a full unfolding of the learning content and more time for discussions and debates, own research through experiments and games. Many parents and teachers share the view that STEM education is only suitable for gifted children. They feel insecure and think that they do not have the necessary training to create a stimulating environment and motivate children to take an interest in STEM issues. This calls for greater collaboration between teachers and parents in STEM education [9].

Another difficulty that teachers face when implementing STEM education is the lack of sufficient resources. Being well resourced with the latest technology and the ability of teachers to use it provides an interactive learning environment for STEM, stimulating creativity and critical thinking in students. What differentiates STEM from traditional education is the integrated learning environment across the four strands (science, technology, engineering, and mathematics) requiring state-of-the-art facilities in line with the continuous advancement of science and technology [17].

The successful implementation of entrepreneurial activities is conditioned by the influence of various factors, among which can be mentioned: the orientation of the educational system towards the preparation of personnel needed by the business and the labor market; the level of introduction of technological innovations and the presence of entrepreneurial traditions and culture [1].

Entrepreneurship has three specific characteristics, namely:

- Orientation towards high success.
- Innovation.
- Setting ambitious goals [8].

R. Neminska considers innovative approaches as a set of interdisciplinary approach, activity-oriented approach, and research approach. They ensure an active cognitive process based on inter-subject connections and productive mental and research activity of students [15].

This is the basic idea behind the concept of entrepreneurship education through STEM. The benefits for young people who have received entrepreneurship education, according to the Entrepreneurship 2020 Action Plan, are limited to learning how to turn ideas into business action [3].

IV. Good practices for applying STEM in entrepreneurship education

The application of STEM in entrepreneurship education is a relatively new teaching method for Bulgaria. In partnership with the Junior Achievement Bulgaria Foundation, activities related to the promotion of practical methods and ideas for entrepreneurship are being implemented. One of the most widespread examples of practically oriented entrepreneurship training is working in a training company. This base used together with STEM is an innovative method that combines an integrated approach, project-based learning, problem-based learning, experiential learning, and research. The training focuses on mastering basic knowledge, skills and attitudes related to entrepreneurship and career guidance. The development and implementation of projects with an entrepreneurial orientation in extracurricular activities and interest clubs, develops the ability to communicate and work in a team, supports students' future career choices. Examples which illustrate this are the educational initiatives carried out in Bulgarian school such as "Your Lesson" project, funded under the Operational Program "Science and Education for Smart Growth"; the National Initiative "Manager for a Day"; National competitions on professions; National competition "Virtual Enterprise"; National competition "Best Training Company of Bulgaria", etc.

In this publication, we share the good experience of the Agricultural Vocational School "Kliment Timiryazev" – town Sandanski, from the application of an interdisciplinary STEM approach to working in a training company.

The conducted research aims to establish the applicability of STEM technology and interdisciplinary learning in extracurricular activities in entrepreneurship. Thirty students of class XI in vocational school took part in the study.

The object of study was the STEM learning conducted through participation in an extracurricular activity of interest, the Beekeeping Club.

The object of study is the impact of applied STEM - science activities on students' activity, motivation, communication, and satisfaction. The following methods were used: analysis of written sources, study of administrative documentation, observation, discussion with students, parents and teachers, statistical methods for processing the obtained data. The proposed methodology was tested with the application of activity-oriented and integrated approach, through the participation of students in projects and interdisciplinary situations with entrepreneurial orientation. The examples include:

- Project on "Is the extinction of bees useful for humans, and should we protect the environment?"

Students seek information from a variety of sources that answer the questions: what is the environment and how can we protect it? What are honey and bee products used for? In the entrepreneurship classes, a business plan is prepared in order to sell the organic bee products and the additional bio cosmetic products such as soaps, creams, toothpaste prepared in the school chemistry lab. Moreover, during the e-trade classes students will acquire skills in launching an e-shop advertising and offering the above-mentioned products. The profit will be donated to local charities. Students develop skills in searching, selecting, and processing information using ICT and prepare a multimedia presentation. Conduct a virtual hive walk, using virtual reality goggles viewing the structure of the hive and the life of a bee colony. Additionally, videos are filmed and then edited and formatted in the ICT lessons, these videos are available to anyone who wishes to learn about this topic. Virtual reality is a new and rapidly growing technological innovation in STEM, available to students in the classroom through free apps. Additionally, they develop their personal and interpersonal skills such as empathy, and social commitment.

- Project on "Solar models/panels/placed in the side extensions of the innovative hive systems

Students construct models that are powered by solar energy through a photovoltaic that delivers the necessary energy to the hive (Fig. 1). In this way, they gain knowledge about alternative energy sources, about sunlight and its possibilities to generate energy. By practicing STEM science activities, learners study the bee colony, combine knowledge of ecology and entrepreneurship, mathematics, and technology, in developing software code and programming microcontrollers and microchips, and this way enhance their educational outcomes.

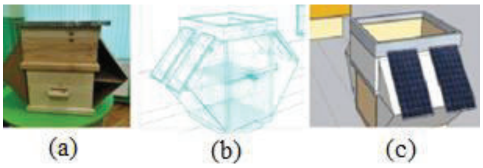


Fig. 1 (a) STEM beehive – complete model, (b) digital model of the STEM beehive, and (c) photovoltaic system

- Interdisciplinary activities on the bee colony structure and the need to construct a hive to optimize the work of the beekeeper

The tasks within the interdisciplinary acuities emphasize on the development of skills for: information search and selection; teamwork and collaboration; the presentation of the constructed prototypes, programming of the sensors and the solar system. Didactic technology involves bringing out the main situational problems of the life of a bee colony, establishing roles and relationships.

Expected outcomes of the topic are the development of analytical and practical-applied skills, through the organization of an interactive learning environment in the constructed stem center at school. Activities are divided into teams to propose solutions. The methods used are: interactive and role-playing games, arranging the wooden cham components /puzzle/ to visualize the model of the stem hive, visualization with poster and group work.

In the process of teamwork, students develop skills in decision-making, sharing, searching for information, presenting their solutions, with new technologies.

Three working teams are formed to play different roles and the training takes the form of a role-play. The class is divided into three working teams (Team 1, Team 2, and Team 3). In each team there are **10** students, divided into 3 groups (A, B and C), who take on different roles.

Team 1 activities - students play the role of the queen bee and her dance as they talk about their lives. A student from the team tells about their activity using their skills to graphically represent the different components of the bee frame in the brood box of the beehive.

Team 2 activity - students take on the role of drones and a student from the team plays the role of a drone speaker and talks about their activity and the importance of protecting the ecosystem.

Team 3 activity - students play the role of bees constructing a STEM hive and those with worker roles talk about their lives as a family. A student from the team shares his activity of programming the micro controllers and processing the data collected from the sensors, emphasizing the importance of power through the solar system.

The results of the study are evaluated according to the following criteria:

Criterion 1 (K1), choice of action option. The indicators for this group are understanding of the situation, reasoning and selection of 1,2 or 3 solution options in the group;

Criterion 2 (K2), cooperativeness and teamwork. The indicators for this group are efficiency of work, organization, good interaction and cooperation, level of self-preparation, expressing and justifying their own opinion and choices, reaching a common solution. The assessment of the learning outcomes of the students included in the groups (A, B and C) is carried out at three levels (high, medium, and low) according to the defined criteria and indicators. Figures 1, 2, and 3 present data on the results of the individual teams. The first team has low work efficiency, is not well-organized and lacks good interaction and cooperation, their self-training is at an average level (Table 1, Fig. 2).

TABLE I
DATA FOR TEAM 1

Criteria	A Group	B Group	C Group
K1	4	0	6
K2	3	2	5

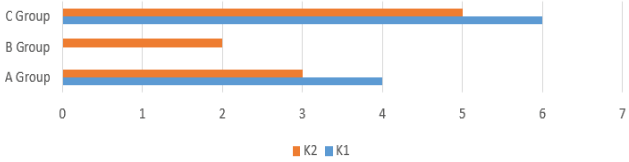


Fig. 2 Graphical representation of Team 1 data

The second team has good work efficiency, well-organized, with good interaction and cooperation, their self-training is at a good level (Table 2, Fig. 3).

TABLE II
DATA FOR TEAM 2

Criteria	A Group	B Group	C Group
K1	5	2	3
K2	3	0	7

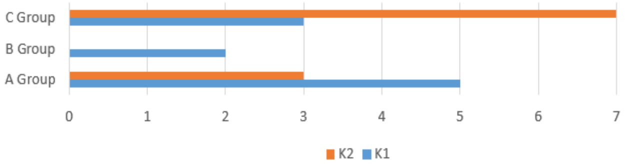


Fig. 3 Graphical representation of Team 2 data

The third team is highly efficient, well-organized, with good interaction and cooperation, their self-training is at a very good level (Table 3, Fig. 4).

TABLE III
DATA FOR TEAM 3

Criteria	A Group	B Group	C Group
K1	5	2	3
K2	2	3	5

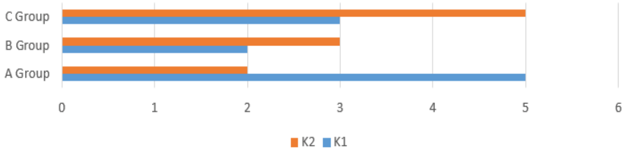


Fig. 4 Graphical representation of Team 3 data

Analysis of the results shows that STEM learning, within a training company, is an innovative technology based on knowledge integration, positively influencing students' cognitive and motivational attitudes. This innovative for Bulgaria approach stimulates, spurs, and motivates students to upgrade knowledge independently and hence to achieve better results. Moreover, to increase active participation of students, it is important to use forms and methods of work that put the student in an active position. Creating a strong link between learning and experience leads to better cognitive outcomes. Project work and activity-learning situations in extracurricular activities of interest, with the application of STEM, meet the approval, of students, parents, and educators.

V. Conclusion

Initiative and entrepreneurship are important personal qualities that every person must possess in order to be successful in the current global economic changes. In order to succeed in a high-tech, information-based society, students must develop their STEM abilities. STEM education is based on self-directed learning, using an interdisciplinary approach. The practice-oriented entrepreneurship education, within a training company, leads to the accumulation of experiences both in and outside the classroom environment, stimulating students to become active, creative, and enterprising citizens. STEM learning within a learning company does not replace subject-specific knowledge but complements, develops and builds on it.

References

[1] D. Doykov, (2007). Cite., p. 61

[2] D. Timmerman, “Project Artifex”, 2017

[3] European Commission, “Entrepreneurship 2020 Action Plan”, INT/679-EESC-2013-941, 22 May 2013

[4] European Commission, “School development and excellent teaching for a great start in life”, Brussels, 30 May 2017, COM/2017/0248 final

[5] European Commission/EACEA/Euridica, “Entrepreneurship Education in European Schools”, 2016, ISBN 978-92-9492-428-5 doi:10.2797/324518 EC-02-16-104-BG-N, pp.8-9 and p.17

[6] I. Keresiev, “Competitive Advantages of Small and Medium Enterprises”, 2017

[7] I. Keresiev, “Professionalizing Management in The Process of Family Business Succession: Securing Sustainable Development of Bulgarian Family Firms”, 2017

[8] I. Keresiev, “Social Entrepreneurship: Nature and Characteristics”, 11 Sept. 2017

[9] I. Mirtschewa, “STEM Education - Main Characteristics and Problems”, Oct. 2021

[10] J. v/d Veer, “EU STEM Coalition”, 11 Dec. 2020

[11] K. Asisyan, “STEM – things we need to know”, Feb. 2022

[12] M. Amato, A. Siri, “Guide to Innovative Pedagogical Approaches in STEM Education”, Florence, Oct. 2019, Erasmus+ Project Number: 2017-1-IT02-KA201-036780

[13] M. Amato, A. Siri, “Guide to Innovative Pedagogical Approaches in STEM Education”, Florence, Oct. 2019, Erasmus+ Project Number: 2017-1-IT02-KA201-036780, p.11

[14] National Assembly of Bulgaria, “The Annual Programme for the participation of the Republic of Bulgaria in the decision-making process of the European Union (EU)”, Jan. 2021, p. 26

[15] R. Neminska, “Research Training in an Academic Environment”, IOSR Journal of Research & Method in Education (IOSR-JRME), 2016

[16] S. Conze, “Achieving a European Education Area by 2025 and resetting education and training for the digital age”, Brussels, Sept. 2020

[17] uchitel.bg, “What do we have to know about STEM”, Mar. 2017

Digital Transformation in Fashion System Design: Tools and Opportunities

M. Tufarelli, L. Giliberti, E. Pucci[1]

Abstract - The fashion industry's interest in virtuality is linked on the one hand to the emotional and immersive possibilities of digital resources and the resulting languages, and on the other to the greater efficiency that can be achieved throughout the value chain. The interaction between digital innovation and deep-rooted manufacturing traditions today translates into a paradigm shift for the entire fashion industry where, for example, the traditional values of industrial secrecy and know-how give way to experimentation in an open as well as participatory way, and the complete emancipation of virtual reality from actual 'reality'.

The contribution aims to investigate the theme of digitisation in the Italian fashion industry, analysing its opportunities and the criticalities that have hindered its diffusion.

There are two reasons why the most common approach in the fashion sector is still analogue:

- the fashion product lives in close contact with the human body, so the sensory perception of materials plays a central role in both the use and the design of the product, but current technology is not able to restore the sense of touch;
- volumes are obtained by stitching flat surfaces that once assembled, given the flexibility of the material, can assume almost infinite configurations. Managing the fit and styling of virtual garments involves a wide range of factors, including mechanical simulation, collision detection and user interface techniques for garment creation.

After briefly reviewing some of the salient historical milestones in the resolution of problems related to the digital simulation of deformable materials and the user interface for the procedures for the realisation of the clothing system, the paper will describe the operation and possibilities offered today by the latest generation of specialised software. Parametric avatars and digital sartorial approach; drawing tools optimised for pattern making; materials both from the point of view of simulated physical behaviour and of aesthetic performance, tools for checking wearability, renderings, but also tools and procedures useful to companies both for dialogue with prototyping software and machinery and for managing the archive and the variants to be made.

The article demonstrates how developments in technology and digital procedures now make it possible to intervene in different stages of design in the fashion industry. An integrated and additive process in which the constructed 3D models are usable both in the prototyping and communication of physical products and in the possible exclusively digital uses of 3D models in the new generation of virtual spaces. Mastering such tools requires the acquisition of specific digital skills and at the same time traditional skills for the design of the clothing system, but the benefits are manifold and applicable to different business dimensions. We are only at the beginning of the global digital transformation: the emergence of new professional figures and design dynamics leaves room for imagination but, in addition to applying digital tools to traditional procedures, traditional fashion know-how needs to be transferred into emerging digital practices to ensure the continuity of the technical-cultural heritage beyond the transformation.

Keywords— Digital fashion, Digital technology and couture, Digital fashion communication, 3D Garment simulation.

[1]M. Tufarelli is Researcher at University of Florence, Italy. DIDA Design Campus (e-mail: margherita.tufarelli@unifi.it).

L. Giliberti is PhD student at University of Florence, Italy. DIDA Design Campus. (e-mail: leonardo.giliberti@unifi.it).

E. Pucci is Junior Researcher at University of Florence, Italy. DIDA Design Campus (e-mail: elena.pucci@unifi.it).

A Data Science Pipeline for Algorithmic Trading: A Comparative Study in Applications to Finance and Cryptoeconomics

Luyao Zhang, Tianyu Wu, Jiayi Li, Carlos-Gustavo Salas-Flores, Saad Lahrichi

Abstract**—**Recent advances in AI have made algorithmic trading a central role in finance. However, current research and applications are disconnected information islands. We propose a generally applicable pipeline for designing, programming, and evaluating algorithmic trading of stock and crypto tokens. Moreover, we provide comparative case studies for four conventional algorithms, including moving average crossover, volume-weighted average price, sentiment analysis, and statistical arbitrage. Our study offers a systematic way to program and compare different trading strategies. Moreover, we implement our algorithms by object-oriented programming in Python3, which serves as open-source software for future academic research and applications.

Keywords**—**Algorithmic trading, AI for finance, fintech, machine learning, moving average crossover, volume weighted average price, sentiment analysis, statistical arbitrage, pair trading, machine learning,

Luyao Zhang is with the Duke Kunshan University, China (e-mail: sunshineluyao@gmail.com).

When Change Is the Only Constant: The Impact of Change Frequency and Diversity on Change

D. Pieters

***Abstract*—** Due to changing societal and economic demands, organizational change has become increasingly prevalent in work life. While a long time change research has focused on the effects of single discrete change-events on different employee outcomes such as job satisfaction and organizational commitment, a nascent research stream has begun to look into the potential cumulative effects of change in the context of continuous intense reforms. This case study of a large Belgian public organization aims to add to this growing literature by examining how the frequency and diversity of past changes impacts employees' appraisals of an newly introduced change. Twelve hundred survey-results were analyzed using standard ordinary least squares regression. Results showed a correlation between high past change frequency and diversity and a negative appraisal of the new change. Implications for practitioners and future research are discussed.

***Keywords*—** change appraisal, change history, organizational changes

Danika Pieters is with the University of Antwerp, Belgium (e-mail: danika.pieters@uantwerpen.be).

Unintended Health Inequity: Using the Relationship Between the Social Determinants of Health and Employer-Sponsored Health Insurance as a Catalyst for Organizational Development and Change

Dinamarie Fonzone

***Abstract*—** Employer-sponsored health insurance (ESI) strategic decision-making processes rely on financial analysis to guide leadership in choosing plans that will produce optimal organizational spending outcomes. These financial decision-making methods have not abated ESI costs. Previously unrecognized external social determinants, the impact on ESI plan spending, and other organizational strategies are emerging and are important considerations for organizational decision-makers and change management practitioners. The purpose of thisstudy is to examine the relationship between the social determinants of health (SDoH), employer-sponsored health insurance (ESI) plans, andthe unintended consequence of health inequity. A quantitative research design using selectemployee records from an existing employer human capital management database will be analyzed. Statistical regressionmethods will be used to study the relationships between certainSDoH (employee income, neighborhood geographic living area, and health care access) and health plan utilization, cost, and chronic disease prevalence. The discussion will include an application of the social gradient of health theory to the study findings, organizational transformation through changes in ESI decision-making mental models, and the connection of ESI health inequity to organizational development and changediversity, equity, and inclusion strategies.

***Keywords*—** employer-sponsored health insurance, social determinants of health, health inequity, mental models, organizational development, organizational change, social gradient of health theory.

Dina Fonzone is with the Cabrini University, United States (e-mail: df10394@cabrini.edu).

A novel Multi-Objective Multi-Mode Multi-Skill mathematical model for Time Cost and Quality Trade-off in Project Scheduling

[A]sina Zabihi[1], [A]nikbakhsh Javadian, [C]alireza Savari Choulabi, [D]mohammadamin Molaei Aloucheh

[a] Department of Industrial Engineering, Mazandaran University of Science & Technology, Babol, Iran; Email: S.zabihi@jpcomplex.com

[a] Department of Industrial Engineering, Mazandaran University of Science & Technology, Babol, Iran; Email: nijavadian@ustmb.ac.ir

[c] Department of Management and Economics, Science and Research Branch, Islamic Azad University, Tehran, Iran;
Email: savarii@jpcomplex.com

[d] Department of Management and Accounting, Shahid Beheshti University, Tehran, Iran;
Email: Mohammadaminmolaei79@gmail.com

Abstract

In this paper, a novel multi-objective multi-mode multi-skilled project scheduling problem with a time cost quality trade-off approach is proposed. In this problem, Each Activity can be performed in different executing modes, and each mode needs different requirements; requirements in this problem are different skills, which should be done by staff. Furthermore, staff members have various skills and specialties and can perform different skills in different activities. To model this problem, an integer linear programming formulation is proposed to optimize concurrently the objectives, including (1) minimizing total project completion time, (2) minimizing total salary of the workforce involved in performing activities, and (3) maximizing quality of workforce on performing skills of activities. Then L_p metric technique is utilized to transform the multi-objective problem to a single objective problem and find the optimal solution; thus, the proposed L_p metric model has been solved optimally by GAMS in small-sized. Analyzing the results showed that the optimal solution obtained from an L_p metric method in small-sized is optimal and reliable. Solving the model has created reasonable and feasible solutions for small-sized problems. The proposed model is firmly NP-hard, and bringing exact solutions is time-consuming.

Keywords: Multi-skilled project scheduling problem; Multi-mode resource constraint project-scheduling problem, Time cost quality trade-off problem; L_p metric,

[1] * Corresponding author; Email: S.zabihi@jpcomplex.com; Tel: +989127923487

1. Introduction

Over the years, project scheduling has been among the most critical areas in operational research applications. During decades from the emergence and application of project management knowledge, its concepts, dimensions, and methods are constantly expanding and improving. These changes seek to bring the proposed models closer to the real problems in today's complex world, which has led to many advances in the science of scheduling optimization.

The resource-constrained project scheduling problem (RCPSP) was first presented by Wiest(1962). RCPSP is one of the most practical problems in operations research. In recent years, significant advances have expanded its dimensions and accurate and innovative solutions. The purpose of presenting the RCPSP is to schedule the activities in the project and determine the appropriate sequence of activities so that the Precedence relations between different activities in the project are fully considered. Various limitations such as constrained resources used in the project Also are fully considered.

Blazewicz et al.(1983) Showed that the RCPSP is an NP-Hard problem. Since the definition of this problem, much research has been done on RCPSP. The study is divided into two parts: 1) presenting various mathematical models for the RCPSP 2) providing different solution methods, including exact, heuristic, and meta-heuristic techniques.

Shou et al.(2015) proposed an integrated PSO algorithm to solve the RCPSP. They offered four different types of answer representations for the PSO algorithm, and for all of them, they used a crossover operator to improve the answers. The computational results showed that the algorithm they developed to solve the RCPSP has a high ability to provide appropriate solutions.

Moukrim et al.(2015) proposed a mathematical model for the RCPSP in the condition of a preemption. They offered a mathematical model for finding the optimal solution to minimize the total project time. They used a branch & price algorithm to solve the real problems presented in the literature, and their proposed algorithm was based on the column generation method. The computational results indicated the proper performance of the algorithm in solving the sample problems presented in the research literature.

Wang et al.(2021) proposed a bi-objective mathematical model to make the resource transfer decisions, aiming to minimize the transfer cost and maximize solution robustness in the presence of activity duration variability because transfers of renewable resources between activities overall incur certain scheduling costs and affect the robustness of a specific schedule in an uncertain environment. They used a novel resource-oriented flow formulation that was different from previous literature. An NSGA-II and a Pareto simulated annealing (PSA) algorithm have been applied as the solution methodologies. Furthermore, *the* ε-constraint method was used to evaluate the effectiveness of the metaheuristics Algorithms. The results indicated that the suggested model and algorithms were applicable and beneficial to the problem in practice.

In the normal state of the RCPSP, it is assumed that each activity can only be performed in a specific mode, while in the real world, each activity can be performed in more than one mode and selectively. Thus, defining a multi-mode resource constraint project-scheduling problem (MMRCPSP) in a project is selecting a mode to perform each activity (executing mode) and determining each activity's start and finish time so that resource constraints and Precedence relations are regarded, and the project’s duration is minimized.

Afshar-Nadjafi et al.(2014) developed a mixed-integer programming model for the MMRCPSP in the condition of the preemption in activities; the objective was to minimize renewable/nonrenewable resource costs and earliness-tardiness costs by a given project deadline and due dates for activities. They proposed a genetic algorithm (GA) to solve the problem. To evaluate the quality of the proposed algorithm, 120 test problems were used. Comparative statistical results reveal that the proposed GA was efficient and effective in terms of the objective function and computational times.

Yuan et al.(2021) formulated a prefabricated building (PB) construction resource-constrained project scheduling with the multi-objective multi-mode optimization model, focusing on the uncertainty of the execution of activity's duration. They used the interval value of the execution time to express it through fuzzy theory and consider the multiple objectives, including time-based profit and cost-based profit. They also proposed a hybrid cooperative co-evolution algorithm (HCOEA) to obtain the highly robust project scheduling and reduce the uncertainty of the execution time of the overall project. They design HCOEA with multi-stage representation for the activity sequencing and the resource allocation, further improving the search efficiency. Finally, benchmarks and datasets with fuzzy processing time were adopted to test their HCOEA. Computational results showed that the HCOEA performed better than the existing state-of-the-art methods.

Time cost and quality are three critical factors in planning and controlling construction projects. The project's time, cost, and quality are determined by the project activities' time, cost, and quality. On the other hand, these criteria for each of the activities in the project network depend on its executing mode. Getting these factors in balance, which minimizes the project duration, the total project cost, and maximizes the total quality, could define the success of a construction project. Usually, using more effective execution techniques will reduce the completion time of an activity, but we will have to pay more for using more efficient resources and technologies. Generally, more inexpensive resources or technologies lead to an increase in the duration of activities and thus an increase in the entire project's duration. Utilizing resources with greater productivity and efficiency or more advanced technologies will save project time and increase the project's direct cost. Simultaneous reduction of time and cost may also lead to reduced quality of projects. Finding an optimal solution to the Time cost quality trade-off problem (TCQTP) includes determining the optimal execution modes for all the activities that make up the project. An optimal combination of time, cost, and quality for all activities is created. As a result, the main target in TCQTP is to allocate limited resources to activities with respecting Precedence relations to achieve the predetermined goals, such as minimizing the project’s completion time, minimizing the project’s total cost, and maximizing the project’s overall quality.

Afruzi et al.(2013) presented a multi-objective mathematical model for TCQTP. They assumed that the resources required for each activity mode were different, and each activity could be considered normal or preemptive. They used a meta-heuristic algorithm called the adjusted fuzzy dominance genetic algorithm to solve the problem. To evaluate the performance of the proposed multi-objective meta-heuristic algorithm, it was compared with four well-known algorithms; NSGAII, NRGA, PAES, MOIWO. The obtained results showed the effectiveness of the proposed algorithm.

Monghasemi et al.(2015) presented a mathematical model combining multi-objective optimization problems and multi-criteria decision-making methods to model the TCQTP. They used Shannon's entropy technique to weigh the objectives of the problem. They used a multi-objective algorithm combining NSGA-II and MOGA to solve the proposed model. The results showed that the integrated algorithm has a lower solution time than the NSGA-II and MOGA algorithms.

Nguyen et al.(2022) Utilized fuzzy logic to model the uncertainty embedding α-cut approach to see the effect of the uncertainty on the time, cost, and quality of the project. Then multi-objective Symbiotic Organism Search (SOS) algorithm was applied to find a set of the optimal solution in different uncertainty levels and provide the project manager several possible actions to implement the project. Some numerical case studies were analyzed to find the model's effectiveness and capability to solve the TCQTP in the construction project. The results illustrated that the proposed model was powerful to find the solution for the shortest project duration with minimum incurred cost and high overall quality in the construction project. Compared to the other widely used methods and other algorithms, the proposed model was proven effective and competitive in solving the TCQTP.

Sharma et al.(2022) proposed an NSGAII based TCQT optimization model for project scheduling. They assumed that each project activity had different alternatives accompanied by other times, costs, and impact on the entire project quality. The main objective was to quantitatively estimate the project's quality and determine the optimal combinations of activities alternatives while minimizing the time cost and maximizing the entire project quality. A pairwise comparison-based analytical hierarchy process (AHP) was employed to determine the relative weight of activities and quality indicators of the project. The developed model was utilized for a case study project; Results of the case study project demonstrated the efficiency of the proposed model in simultaneous optimization of time, cost, and quality of the project.

2. Multi-skill project scheduling problem

The Multi-Skill Project Scheduling problem (MSPSP) was first proposed by Néron(2002). He combined the classic RCPSP with the Multi-Purpose Machine (MPM) model to provide a model that minimizes project completion time.

MSPSP is an extended model of the MMRCPSP problem. The main difference between the MSPSP problem and the RCPSP problem or the MMRCPSP problem is considered in the resources and the method of defining the requirement of each resource. In this problem, the resources are the staff members assigned to the project. The method of using these resources is determined based on the skills of each of them during the project. Therefore, in this type of problem, each human resource has a set of skills, and each activity needs a set of skills to be performed. In MSPSP, the skills of staff and the needs of each activity to different skills are defined and specified. By this definition, staff members assigned to each activity are a subset of individuals who can meet the skill needs of that activity. The large number of modes for each activity to be performed is rooted due to the ability of each member to do different skills in activities. If in MSPSP all members (resources) have only one skill, we are dealing with the same classic RCPSP model. MSPSP is the case for various disciplines, including construction, IT projects, healthcare, and process systems.

Montoya et al.(2014) proposed a mathematical model that combines the RCPSP problem with the MPM problem. The purpose of presenting this model was to find the optimal schedule of activities to minimize the completion time of the whole project. They assumed that the resources used in this project were human resources with different skills. Therefore, to perform any required skill in each activity, some staff must be allocated to that activity. They proposed a new method that combines the column generation approach with the Branch & Price algorithm to solve this problem. The computational results showed that the new method used to solve the proposed model could obtain the optimal global solution of the model in small and medium sizes in a reasonable time.

Shahnazari-Shahrezaei et al.(2017) proposed a particular type of MSPSP called Multi-Objective Multi-Skilled Project Scheduling Problem (MOMSPSP), which incorporated some new objectives in the MSPSP and developed a multi-objective mixed-integer non-linear programming (MINLP) model. The model was exactly solved for small-sized instances using a CPLEX solver. Differential Evolution (DE) and Particle Swarm Optimization (PSO) based meta-heuristic algorithms were utilized to solve such an NP-hard problem for medium and large-sized instances. To evaluate the performance of the propounded algorithms, results were compared with each other and the optimal solution obtained by the CPLEX solver for small-sized instances. Finally, the designed DE algorithm was explored as the superior proposed algorithm for solving the proposed MOMSPSP in terms of some performance metrics.

Zabihi et al.(2019) proposed a novel multi-objective multi-skilled project scheduling problem that the aims of the model were minimizing the total salary of the workforce involved in performing activities, minimizing total completion time of activities, and maximizing the workforce's efficiency on performing skills of activities by non-linear learning curve function. They proposed two hybrid TLBO based algorithms inspired by the multi-objective concept of the MOPSO and MOIWO algorithms to solve the model. The results showed that the MOPTLBO algorithm had better performance in terms of the MID, while the NSGA-II algorithm was the best in terms of the SNS metric. Furthermore, the FSTLBO algorithm was best for the MS, and Spacing mean. The results showed that the proposed hybrid FSTLBO algorithm had a better overall performance on solving various instances.

Polo□Mejía et al.(2021) dealt with a new variant of MSPSP with respecting partial preemption, in which only a subset of resources can be released during the preemption periods. They proposed a series of heuristic algorithms to solve instances arising from an industrial application. Therefore, they presented a serial greedy algorithm based on priority rules and a flow problem for resource allocation. They introduced a binary tree-based search algorithm and a greedy randomized adaptive search procedure (GRASP) to improve the solutions of the greedy algorithm. Finally, they proposed an extensive neighborhood search (LNS) algorithm integrating exact and heuristic methods. The best solution quality and execution time results were obtained by combining the GRASP algorithm and LNS approach.

Afshar-Nadjafi(2021) reviewed and surveyed the literature on scheduling problems under multi-skilled and flexible resources. The paper's primary purpose was to help researchers and scholars enter the multi-skilling experience with a comprehensive overview of existing models and methods and identify new research directions. He reviewed and classified 160 articles published from 2000 to 2020 based on the objective functions, the mathematical modeling, the solving methodologies, and the potential applications. The results showed that the main focus of the existing research in this field had been allocated to project scheduling problems (53.12%), mixed integer programming models (54.2%) and meta-heuristics (28.7%) as solving method, cost (39.4%) as a single objective model (68.6%), and deterministic condition for parameters (85.5%) on the top. It also showed that 68.8% of research regarded a single objective, of which 13.8% and 17.6% of the study have developed models with two and more objectives, respectively.

After reviewing research in RCPSP, MMRCPSP, TCQTP, and MSPSP, a novel mathematical model for combining TCQTP, MMRCPSP, and MSPSP when project dealing only with staff resources will be presented, which in turn is new.

3. Model description

This section first presents and formulates the mathematical model for MOMMMSPSP, indices, parameters, and decision variables. Afterward, a mathematical model for the related MOMMMSPSP regarding the following assumptions is defined.

Herein, a novel three objective model is proposed to:

- Determine workforce assigned to skills of each activity
- Determine the best possible starting times of all activities
- Determine the best mode of each activity

Herein, an integer linear programming formulation is presented to formulate the problem. The proposed model aimed to optimize three different conflicting objectives, including 1) minimizing the project’s total completion time, 2) maximizing the project’s total quality, and 3) minimizing the project’s total cost. In the following, the major underlying assumptions of the model, indices, parameters, decision variables, and the mathematical model are presented.

3.1 The most significant assumptions:

- The project's network is defined as an AON network.
- Activities are indexed in a topological form.
- All the required resources are workforce and are always available.
- Performing each activity needs several skills.
- Each workforce has various skills (either has a specialty or does not)
- All the execution times are considered integer.
- Each activity execution's time is an integer, depending on its selective executing mode.
- Activities' setup times and time of assigning resources to the activities are ignorable.
- Interrupting activities during their execution is not allowed (preemption is not allowed).
- Each activity must be performed under one mode during its execution from its potential executing modes, and each mode has its unique duration time and requirement.
- Each workforce performs each skill by defined cost and quality.
- Each workforce cannot perform more than one activity simultaneously.
- Each workforce can be allocated to different activities if those activities are not executed concurrently.
- All allocated staff to different skills in an activity must start and finish that skills simultaneously.
- Performing skills of each activity need several predefined numbers of workforces (depending on its executing mode).
- The workforce allocated to each activity should be assigned to particular predefined skills, and their job should be finished in a specific time.
- The project's total quality is calculated as the weighted average quality of all the staff involved in performing the project's different skills.
- All the staff salary via their assignment in activities calculates the project's total cost.

3.2. Notations

Indices and sets

n The index of the last activity

A_1 The start node of the project

A_n The end node of the project

K number of skills.

M number of executing mode.

S number of staff members.

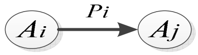
A_i , $i\in\{1..., n\}$ set of not preemptive activities of the project.

E_k , $k\in\{1..., K\}$ set of skills.

O_m , $m\in\{1..., M\}$ set of modes for performing activities.

R_s , $s\in\{1... S\}$ set of staff members.

G: the project graph defined by (A: activity, E: precedence relation, P: duration).

$(A_i.A_j) \in E$ If there exists a precedence relation between A_i andA_j ; 

Parameters

P_{im}: Duration of activity A_i in mode m, $i \in \{1..., n\}$ and m∈*{1..., M}*.

$t \in \{0, 1... Tmax\}$: Starting time of project's activities; $T_{max} = \sum_{i=1}^{n} max\ (p_{i1}.p_{i2}.....p_{iM})$

w_i: Weight of activity A_i ,$\sum_{i=1}^{n} w_i = 1$.

C_{sk}: Salary of staff *Rs* for executing skill E_k per day.

q_{sk}: Quality of performing skill E_k by staff *Rs* in the project; $[0, 1]$.

b_{imk}: number of required resources for performing skill the E_k during the execution of activity A_i in mode O_m.

r_{sk} =1, if staff member *Rs* can perform skill E_k; 0, otherwise.

Decision Binary Variables

v_{im} 1, if activity $\underline{i}$ is done in executing mode $\underline{m}$; 0 otherwise.

x_{imst} 1, if staff member $\underline{s}$ begins to work for activity $\underline{i}$ in mode $\underline{m}$ at time $\underline{t}$; 0, otherwise.

y_{imsk} 1, if staff member $\underline{s}$ assigned to activity $\underline{i}$ in mode $\underline{m}$ at skill $\underline{k}$; 0, otherwise.

3.3. The proposed mathematical formulation

The developed MOMMMSPSP is now formulated as a multi-objective integer linear programming model:

$$Z_1 = \text{Min} \sum_{m=1}^{M} \left[\frac{\sum_{s=1}^{S} \sum_{t=0}^{T\max} x_{nmst} * (t + p_{nm})}{\sum_{k=1}^{K} b_{nmk}} \right] \tag{1}$$

$$Z_2 = \text{Min} \left\{ \sum_{i=1}^{n} \left[\sum_{m=1}^{M} \sum_{k=1}^{K} \sum_{s=1}^{S} (y_{imsk} * C_{sk} * p_{im}) \right] \right\} \tag{2}$$

$$Z_3 = Max \sum_{i=1}^{n} \mathrm{W}_i * \left[\sum_{m=1}^{M} \left[\frac{\sum_{k=1}^{K} \sum_{s=1}^{S} y_{imsk} * q_{sk}}{\sum_{k=1}^{K} b_{imk}} \right] \right] \tag{3}$$

Subject to:

$$\sum_{m=1}^{M} v_{im} = 1; \forall (i) \tag{4}$$

$$\sum_{m=1}^{M} \frac{\sum_{s=1}^{S} \sum_{t=0}^{T\max} x_{imst} * (t + p_{im})}{\sum_{k=1}^{K} b_{imk}} \leq \sum_{m'=1}^{M} \frac{\sum_{s'=1}^{S} \sum_{t'=0}^{T\max} x_{jm's't'} * t'}{\sum_{k'=1}^{K} b_{jm'k'}} ; \forall (i, j) \tag{5}$$

$$x_{imst} \leq v_{im} ; \forall (i, m, s, t) \tag{6}$$

$$y_{imsk} \leq v_{im} ; \forall (i, m, s, k) \tag{7}$$

$$\sum_{m=1}^{M} \sum_{t=0}^{T\max} x_{imst} \leq 1; \forall \ (i, s) \tag{8}$$

$$\sum_{t=0}^{T\max} x_{imst} * t \le \frac{\sum_{h=1}^{S}\sum_{t=0}^{T\max} x_{imht} * t}{\sum_{k=1}^{K} b_{imk}} ; \forall (i, s, m) \tag{9}$$

$$y_{imsk} \le r_{sk} ; \forall (i, m, s, k) \tag{10}$$

$$\sum_{s=1}^{S}\sum_{t=0}^{T\max} x_{imst} = (\sum_{k=1}^{K} b_{imk}) * v_{im} ; \forall (i, m) \tag{11}$$

$$\sum_{s=1}^{S} y_{imsk} = (b_{imk}) * v_{im} ; \forall (i, m, k) \tag{12}$$

$$\sum_{m=1}^{M}\sum_{i=1}^{n}\sum_{d=t-p_{im}+1}^{t} x_{imsd} \le 1; \forall (s, t \in Tmax) \tag{13}$$

$$\sum_{t=0}^{T\max}\sum_{m=1}^{M} x_{imst} = \sum_{k=1}^{K}\sum_{m=1}^{M} y_{imsk} ; \forall (i, s) \tag{14}$$

$$\sum_{m=1}^{M} y_{imsk} \le 1; \forall (i, s, k) \tag{15}$$

$$\sum_{m=1}^{M} x_{imst} \le 1; \forall (i, s, t) \tag{16}$$

$$x_{imst}, y_{imsk}, v_{im} \in \{0,1\}; \quad \forall (i, m, s, t, k) \tag{17}$$

The first objective function in Equation (1) minimizes the project's total completion time. The second objective function in Equation (2) minimizes the total amount of salaries that should be paid for the workforce involved in performing the project's various skills in activities. The third objective function in Equation (3) aimed to maximize the project's total quality, defined as the weighted average quality of the workforces allocated to the activities' different skills. Equation (4) ensures that each project activity can only be executed in one mode. Equation (5) considers precedence relations (activity *i* is the precedent for activity *j*). Equation (6) guarantees this principle that staff *s* at time *t* can start working on activity *i* in mode *m* provided that activity *i* is specified to perform in mode *m*. Equation (7) enforces that Staff *s* with skill *k* can be assigned to activity *i* in mode *m* provided that activity *i* is determined to perform in mode *m*. Equation (8) guarantees that one staff member can be allocated to a maximum of one activity at a time *t* in the project. Inequality (9) ensures that the entire workforce given for various skills of each activity should concurrently start their job. Equation (10) implies that each workforce assigned to each skill of the activity must be able to perform that skill. Equation (11) guarantees that the total number of staff members assigned to each activity equals the number of staff members with diverse specialties at different skills required for executing that activity. Equation (12) ensures that the total number of staff members at a specific skill allocated to one activity must be equal to the number of staff members required at the considered skill of that activity. Equation (13) guarantees the non-preemptive assignment of workforces to the skills of each activity. In other words, the workforce assigned to a predefined skill cannot be released until the skill is thoroughly performed. Equality (14) describes that if a staff member is assigned to an activity, they should start working on it at only a one-time and must be allocated to one required skill. Equation (15) ensures that any staff with any skill in any activity can be assigned in a maximum of one executive mode. Equation (16) provides that any staff member can start work from the project time horizon in any activity in a maximum of one executable mode. Set of Constraint (17) determines that all decision variables used in the model are binary.

4. Solution procedure

One of the most valuable techniques for dealing with multi-objective optimization problems is transforming a multi-objective model into a single one. Since the MOMMMSPSP model is a multi-objective, integer linear programming model where objective functions are entirely inconsistent, we use Deb(2014) L_p *metric* method. According to this method, a multi-objective problem is solved by respecting each objective function separately. Then a single objective is reformulated, which aims to minimize the sum over the normalized difference between each objective and corresponding optimum value.

Our proposed model assumes that three objective functions are Z_1, Z_2, and Z_3. Based on the L_p metric method, the model should be solved separately for these three objective functions. Then the optimal values for these three objectives are Z_1^*, Z_2^*, Z_3^*. Now, the L_p *metric's* objective function can be formulated as Equation (18):

$$LP = \text{Min}\,Z_4 = \left\{ \left[\frac{Z_1^* - Z_1}{Z_1^*}\right]^2 + \left[\frac{Z_2^* - Z_2}{Z_2^*}\right]^2 + \left[\frac{Z_3^* - Z_3}{Z_3^*}\right]^2 \right\}^{\frac{1}{2}} \tag{18}$$

Using L_p *metric,* objective function, and respecting the MOMMMSPSP model's constraints, we transformed the three objective functions given in Equation (1) – (3) into a single Equation (18) which is integer non-linear programming model; therefore, GAMS software (BARON solver) can efficiently solve this integer non-linear programming, and the optimum point will be gain.

5. Computational results

5.1. Case description (Illustrative numerical sample)

GAMS software was run on a PC with Intel_Core i7 CPU with 8 GB RAM to solve a numerical small-sized sample to verify the proposed mathematical model and illustrate the model's performance. The sample was solved in a computational time equal to 6 minutes and 25 seconds. The sample includes six activities, four staff members, three specialties, and three executing modes. In Figure 1, the precedence network of the project has been depicted. Table 1 shows the number of staff members with specific specialties required for executing each activity at each mode. For example, activity A_3 needs two staff members with skill_3 and one with skill_1 in mode_3 with a processing time of 8 days.

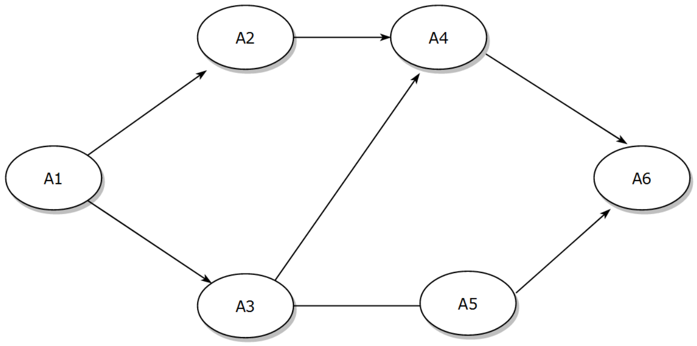


Figure 1. Precedence network of the illustrative example

TABLE 1
The number of staff members with different specialties required for executing each activity to varying skills in each mode and weight of each activity for the illustrative example

Activity	predecessor	W_i	O_1				O_2				O_3			
			p_{im}	E_1	E_2	E_3	p_{im}	E_1	E_2	E_3	p_{im}	E_1	E_2	E_3
A_1	_	0.25	10	0	1	0	7	0	1	1	5	1	1	1

A_2	A_1	0.20	5	1	0	0	4	1	0	1	3	1	1	1
A_3	A_1	0.20	15	1	0	1	12	1	1	1	8	1	0	2
A_4	A_2, A_3	0.15	5	0	0	1	4	0	0	1	2	0	1	1
A_5	A_3	0.10	10	1	0	1	8	1	0	1	4	1	1	1
A_6	A_4, A_5	0.10	5	0	0	1	3	0	0	1	3	0	1	1

In Table 2, the capability of each staff member to do skills is given. Table 3 shows the quality of performing each skill by each staff member in the project. Table 4 illustrates the salary cost of executing skills by staff members in each activity per day.

TABLE 2
The ability of each staff member to do a specialty in the illustrative example

r (s, k)	E_1	E_2	E_3
R_1	1	0	1
R_2	1	1	0
R_3	1	0	1
R_4	0	1	1

TABLE 3
Quality of doing each skill by each staff in each activity

q (s, k)	E_1	E_2	E_3
R_1	0.9	0	0.5
R_2	0.7	0.6	0
R_3	0.9	0	0.5
R_4	0	0.7	0.5

TABLE 4
Salary of performing skills by each staff per working day

c (s, k)	E_1	E_2	E_3
R_1	5	0	3
R_2	5	4	0
R_3	5	0	3
R_4	0	5	4

5.2. Computational Results of multi-objective solution and L_p metric method

In table5, the presented multi-objective model for the MOMMMSPSP is solved by GAMS software, and Optimal values for these three objective functions are obtained as Z_1^*, Z_2^*, Z_3^*.

Table 5
The values obtained for three objective functions by GAMS

Z_1^*	20
Z_2^*	224
Z_3^*	0.715

According to the L_p *metric* method, a multi-objective problem converts to a single objective problem, aiming to minimize the sum over the normalized difference between each objective and corresponding optimal value. In table6, the L_p *metric* transformed objective function considering the MOMMMSPSP model's constraints are solved by GAMS, and values for variables are shown. For example, activity A_5 was executed in mode O_3 in four days, staff R_1 performed skill E_1, staff R_2 performed skill E_2, and staff R_4 performed skill E_3.

Table 6
The variables values obtained via single objective functions by L_p metric

Activity	Executing mode	start	Assigned staffs			
			R_1	R_2	R_3	R_4
A_1	O_3	0	E_3	E_2	E_1	-
A_2	O_1	5	-	-	E_1	-
A_3	O_3	5	E_3	E_1	-	E_3
A_4	O_2	13	-	-	E_3	-
A_5	O_3	13	E_1	E_2	-	E_3
A_6	O_2	17	E_3	-	-	-

In Figure 2, the obtained scheduling solution, including the assignment of each staff member to each activity at each skill, starting and finishing times of each activity is represented.

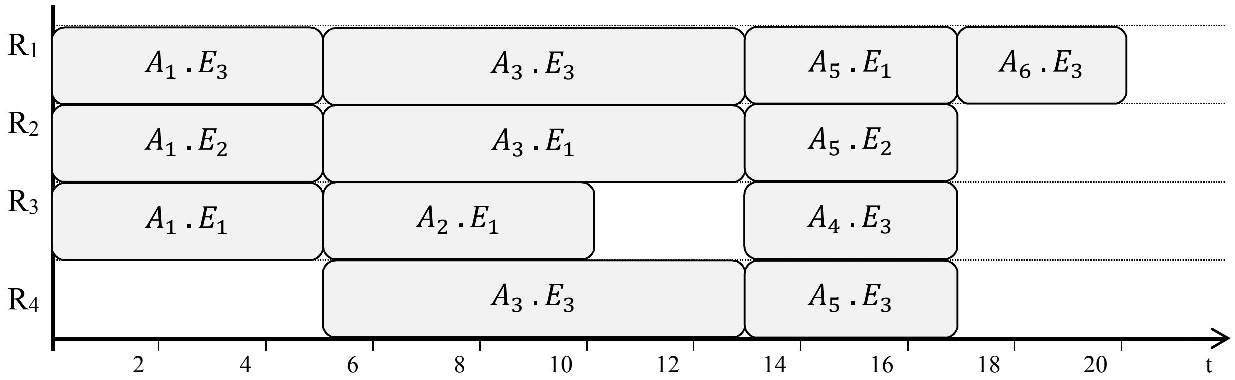


Figure 2. Gantt chart of staff member assignment and scheduling activities for the illustrative example

As seen in the Gantt chart depicted in Figure 2, activity A_1 started at time 0 and finished at time 5, three staff members, including R_1 assigned to skill E_3 , R_2 assigned to skill E_2, and R_3 assigned to skill E_1. Similarly, the assignment of the other staff members can be interpreted. Considering the staff members' assignment due to executing the project activities as depicted in Figure 2, it could be concluded that all model constraints, including Equations (4)-(17), are respected. In addition, given the optimal scheduling solution obtained by the GAMS and its display in the Gantt chart, it is perceived that the assignment of staff members to project activities exactly follows the problem assumptions with no violation of model constraints.

6. Conclusion

A novel multi-objective multi-mode multi-skilled project scheduling problem with a time cost quality trade-off approach is proposed in this paper. An integer linear programming formulation is propounded for modeling the problem to concurrently optimize the objectives, including (1) minimizing total project completion time, (2) minimizing total salary of the workforce involved in performing activities, and (3) maximizing quality of workforce on performing skills of activities. Then L_p *metric* technique is utilized to find the optimal solution; thus, the proposed model has been solved optimally by GAMS in small-sized. Analyzing the results achieved solving the problem in different sizes draw the following conclusions :

1) Solving the developed mathematical model has created reasonable and feasible solutions for small-sized problems .

2) The optimal solution obtained from solving a mathematical model that was transformed to an L_p *metric* method in small-sized is optimal, reliable also proposed model is firmly NP-hard, and bringing exact solutions is time-consuming.

A future extension of this paper is to:

- A researcher can utilize efficient lower bounds resulting from exact methods for solving the developed MOMMMSPSP, leading to better solutions.

- A researcher can apply the presented model for a real situation by utilizing developed heuristic and meta-heuristic algorithms to solve the problem.

- A researcher can consider the model's parameters and variables as fuzzy numbers and develop the model.

References

Afruzi, E. N., Roghanian, E., Najafi, A., & Mazinani, M. (2013). A multi-mode resource-constrained discrete time–cost tradeoff problem solving using an adjusted fuzzy dominance genetic algorithm. *Scientia Iranica, 20*(3), 931-944.

Afshar-Nadjafi, B. (2021). Multi-skilling in scheduling problems: A review on models, methods, and applications. *Computers & Industrial Engineering, 151*, 107004.

Afshar-Nadjafi, B., & Arani, M. (2014). Multimode preemptive resource investment problem subject to due dates for activities: formulation and solution procedure. *Advances in operations research, 2014*.

Blazewicz, J., Lenstra, J. K., & Kan, A. R. (1983). Scheduling subject to resource constraints: classification and complexity. *Discrete applied mathematics, 5*(1), 11-24.

Deb, K. (2014). Multi-objective optimization *Search methodologies* (pp. 403-449): Springer.

Monghasemi, S., Nikoo, M. R., Fasaee, M. A. K., & Adamowski, J. (2015). A novel multi-criteria decision-making model for optimizing time–cost–quality trade-off problems in construction projects. *Expert systems with applications, 42*(6), 3089-3104.

Montoya, C., Bellenguez-Morineau, O., Pinson, E., & Rivreau, D. (2014). Branch-and-price approach for the multi-skill project scheduling problem. *Optimization Letters, 8*(5), 1721-1734.

Moukrim, A., Quilliot, A., & Toussaint, H. (2015). An effective branch-and-price algorithm for the preemptive resource-constrained project scheduling problem based on minimal interval order enumeration. *European Journal of Operational Research, 244*(2), 360-368.

Néron, E. (2002). *Lower bounds for the multi-skill project scheduling problem.* Paper presented at the Proceeding of the eighth international workshop on project management and scheduling.

Nguyen, D.-T., Le-Hoai, L., Tarigan, P. B., & Tran, D.-H. (2022). Tradeoff time cost quality in repetitive construction project using fuzzy logic approach and symbiotic organism search algorithm. *Alexandria Engineering Journal, 61*(2), 1499-1518.

Polo□Mejía, O., Artigues, C., Lopez, P., Mönch, L., & Basini, V. (2021). Heuristic and metaheuristic methods for the multi□skill project scheduling problem with partial preemption. *International Transactions in Operational Research*.

Shahnazari-Shahrezaei, P., Zabihi, S., & Kia, R. (2017). Solving a multi-objective mathematical model for a multi-skilled project scheduling problem by particle swarm optimization and differential evolution algorithms. *Industrial Engineering & Management Systems, 16*(3), 288-306.

Sharma, K., & Trivedi, M. K. (2022). AHP and NSGA-II-Based Time–Cost–Quality Trade-Off Optimization Model for Construction Projects *Artificial Intelligence and Sustainable Computing* (pp. 45-63): Springer.

Shou, Y., Li, Y., & Lai, C. (2015). Hybrid particle swarm optimization for preemptive resource-constrained project scheduling. *Neurocomputing, 148*, 122-128.

Wang, J., Hu, X., Demeulemeester, E., & Zhao, Y. (2021). A bi-objective robust resource allocation model for the RCPSP considering resource transfer costs. *International Journal of Production Research, 59*(2), 367-387.

Wiest, J. D. (1962). The scheduling of large projects with limited resources: carnegie inst of tech pittsburgh pa graduate school of industrial administration.

Yuan, Y., Ye, S., Lin, L., & Gen, M. (2021). Multi-objective multi-mode resource-constrained project scheduling with fuzzy activity durations in prefabricated building construction. *Computers & Industrial Engineering, 158*, 107316.

Zabihi, S., Kahag, M. R., Maghsoudlou, H., & Afshar-Nadjafi, B. (2019). Multi-objective teaching-learning-based meta-heuristic algorithms to solve multi-skilled project scheduling problems. *Computers & Industrial Engineering, 136*, 195-211.

Development and Validation of Integrated Continuous Improvement Framework for Competitiveness: mixed Research of Ethiopian Manufacturing Industries

Haftu Hailu Berhe, Hailekiros Sibhato Gebremichael, Kinfe Tsegay Beyene, Haileselassie Mehari

Haftu Berhe is with the Ethiopian Institute of Technology, Ethiopia (e-mail: bhaftuh@gmail.com).

Abstract - The purpose of the study is to develop and validate integrated literature-based JIT, TQM, TPM, SCM and LSS framework through a combination of the PDCA cycle and DMAIC methodology. The study adopted a mixed research approach. Accordingly, the qualitative study employed to develop the framework is based on identifying the uniqueness and common practices of JIT, TQM, TPM, SCM and LSS initiatives, the existing practice of the integration, identifying the existing gaps in the framework and practices, developing new integrated JIT, TQM, TPM, SCM and LSS practice framework. Previous very few studies of the uniqueness and common practices of the five initiatives are preserved. Whereas the quantitative study working to validate the framework is based on empirical analysis of the self-administered questionnaire using a statistical package for social science. A combination of the PDCA cycle and DMAIC methodology stand integrated CI framework is developed. The proposed framework is constructed as a project-based framework with five detailed implementation phases. Besides, the empirical analysis demonstrated that the proposed framework is valuable if adopted and implemented correctly. So far, there is no study proposed & validated the integrated CI framework within the scope of the study. Therefore, this is the earliest study that proposed and validated the framework for manufacturing industries. The proposed framework is applicable to manufacturing industries and can assist in achieving competitive advantages when the manufacturing industries, institutions and government offer unconditional efforts in implementing the full contents of the framework.

Key words - integrated continuous improvement framework, just in time, total quality management, total productive maintenance, supply chain management, lean six sigma

INTRODUCTION

Continuous Improvement (CI) has long been adopted by many successful companies just to drive out wastes and variations so as to enhance process performance. The literature on this concept becomes so crowded with large numbers of definitions. These definitions are categorizing into two, broader and narrower. The broader definition of CI, it encompasses production and quality management methodologies to achieve organizational excellence [1], [2]. Second is the narrower definition, which is an improvement of the workplace (“gemba”) derived based on the proposals from the workers on the basis of a QCC and a suggestion system [3].

Hence, for the purpose of this study, the broader definition of CI is adapted and refers to Just in Time (JIT), Total Quality Management (TQM), Total Productive Maintenance (TPM), Supply Chain Management (SCM), and Lean Six Sigma (LSS). For example, JIT concentrates on continuously reducing and ultimately eliminating all forms of waste through maintaining speed of producing and delivering the right parts, in the right amount, at the right time using the minimum necessary resources [4]. TQM primarily addresses product improvement customers in a supply chain process [7]; LSS is a business strategy and methodology that increases process performance resulting in enhanced customer satisfaction and improved bottom-line results [8].

So that, from this perspective, synthesizing these CI initiatives into an integrated framework results in the combination of the best of the five programs as the latest generation improvement system which integrates human and technical elements into a programme that links and sequences improvement tools to an overall approach for industrial performance and business growth [9]. This integration is known as Integrated Continuous Improvement Framework (ICIF) with the category of implementation framework.

Despite few ICIF have been proposed with the trend of Plan-Do-Check-Act (PDCA) cycle , for example, integration of JIT and TQM [5], [10], [11]); integration of JIT, TQM, TPM and SCM [9]; integration of Lean and Six Sigma [12], [13]; integration of JIT, TQM and SCM [7]; integration of LSS, Six Sigma and TQM [13]; integration of JIT, TQM and TPM [4]; integration of TPS, TQM and TPM [14]; integration of JIT, TQM and LSS [15], several authors [4], [13], [14]) argue that there is very limited literature on developing and validating integration of CI methods with methods in relation to a specific framework. The researchers also agree with authors and even there is no previous study proposing and validating ICIF for joint implementation of JIT, TQM, TPM, SCM and LSS practices with the trend of combination of PDCA cycle and DMAIC (Define, Measure, and customer satisfaction through application of quality principles to all facets of an organization [5]. TPM principally emphasizes improvement of the production facility, machinery, and equipment [6]. SCM refers to the management of a network of interconnected businesses process involved in the provision of product and service packages required by the end Analyze, Improve and Control) methodology. Thus, this research aims and attempted to fill this gap by developing and validating the integrated CI framework with the focus of JIT, TQM, TPM, SCM and LSS practices which is vital for manufacturing industries excellence.

The researchers have, therefore, focused on manufacturing industries with a view to determine the following two research questions: RQ1. What is the integrated continuous improvement framework that can be used to enhance the competitiveness of manufacturing industries? RQ2. How is the perceived level of the practicability of the proposed framework in the Ethiopian manufacturing industries?

Based on it, this research will present developed implementation framework and empirical evidence showed the practicability of developed implementation framework. Undoubtedly, this research contributes an inclusive review and empirical evidence on continuous improvement approaches principally concentrating on JIT, TQM, TPM, SCM, LSS, PDCA cycle, DMAIC methodology, relationships and integration systems of the five initiatives. The research in keeping with detecting both theoretical and empirical literature gaps.

I. METHODOLOGY – RESEARCH FRAMEWORK, DESIGN AND APPROACH

A. Research Framework

As stated in the introduction, this study interests on developing and validating integrated

CI framework with the focus of JIT, TQM, TPM, SCM and LSS practices from various previous studies. Based on conducting extensive literature review, the study identified and adapted unique practices, common success factors, implementation procedure and potential motivations from [4], [5], [7], [10], [11], [13], [14], [15], [16], [17], [18]. Thus, the study developed conceptual framework (Figure 1) to demonstrate the application of integrated continuous improvement framework in regarding to unique practices, common success factors and implementation procedure that enable in achieving manufacturing industries competitiveness in relation to potential motivations of operational, innovation and business results.

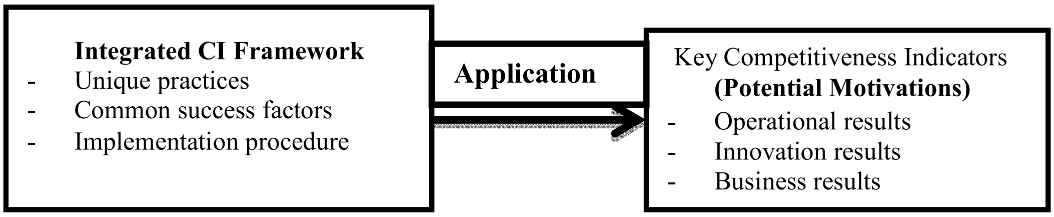


Fig.1 Conceptual framework (the researchers)

B. Research Design

The study adopted both an inductive and deductive approach, and its goal is to report how competitiveness could be achieved through implementation of an integrated framework which develops based on qualitative research and validated through quantitative research. The manufacturing industries currently implementing CI initiatives are considered as the population of this study. The sampling selection of this research is non-probability sampling and purposive sampling. The managers from multi-departments of manufacturing industries including Quality, Production/Operations, Product Development Center, Supply & Procurement, Sales & Marketing, Human Resources and Finance are the target population that has been chosen for the questionnaire survey. The numbers of participants involved in this study is fair-enough to provide adequate feedback to develop and validate the proposed framework, where the percentage of the participants in the questionnaire is relatively high and acceptable [19]. Since the study is targeted on investigation of the practicability of the integrated CI framework, then collection of primary data were obligatory. Accordingly, data were collected through self-administered questionnaire which contains eight questions in relation to the proposed framework and twenty one questions in regarding to the stages of the proposed framework (i.e. implementation procedure).

C. Research Approach

The approach of this research follows seven ways in addressing development and validation of integrated continuous improvement framework (Figure 2). The first is explanation of problems relating to the existing integrated CI framework and the methodology used to implement the integrated framework, and consequently stating the research questions and main research purposes; the second is conducting the literature review and it explains the CI programs, unique and common factors of CI programs; the third approach is identifying and explaining the existing integration approach and gaps in all aspects of the integrated CI practices; the fourth approach is development of the ground-breaking integrated framework; the fifth approach is validation of the proposed framework (FW) by employing reliability and validity analysis, and descriptive analysis of the proposed framework and implementation procedure; the sixth approach is describing the conclusion and recommendations based on the findings of this study and similarly the final (seventh) approach is describing the implications of the developed framework in theoretical and managerial perspectives. Thus, the study is mixed, both quantitative and qualitative pattern analysis to develop and validate the framework which is potential opportunity for manufacturing industries performance improvement. Moreover, the study has presented the analytical results in the forms of tables and charts.

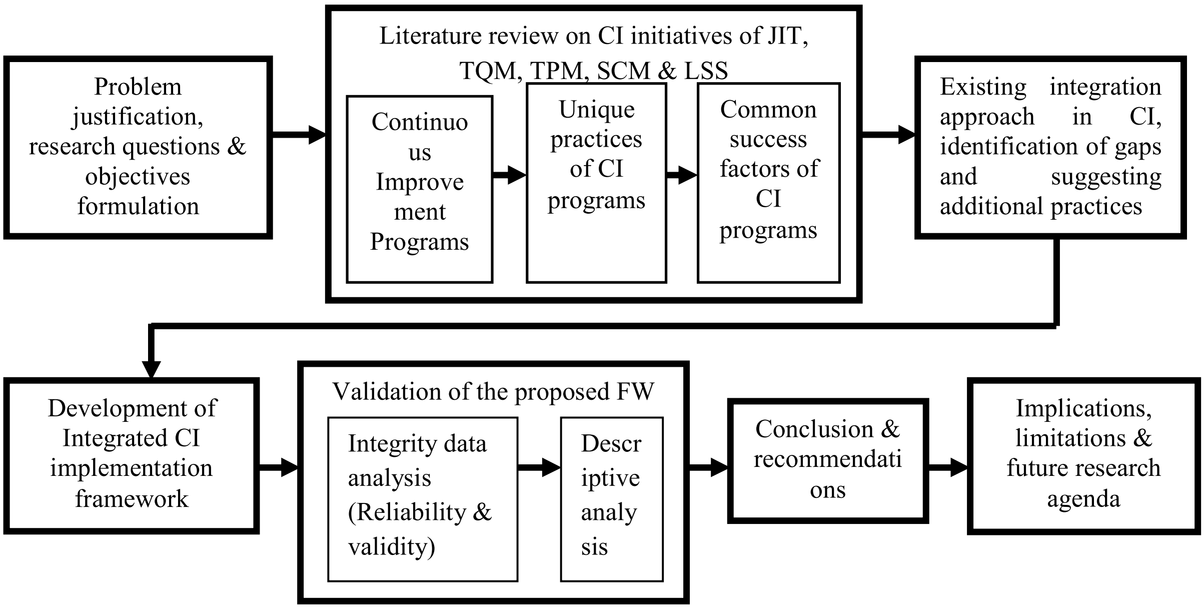

Problem justification, research questions & objectives formulation
Literature review on CI initiatives of JIT, TQM, TPM, SCM & LSS
Continuous Improvement Programs
Unique practices of CI programs
Common success factors of CI programs
Existing integration approach in CI, identification of gaps and suggesting additional practices
Development of Integrated CI implementation framework
Validation of the proposed FW
Integrity data analysis (Reliability & validity)
Descriptive analysis
Conclusion & recommendations
Implications, limitations & future research agenda

Fig. 2 Research methodology framework (the researchers)

II. THE LITERATURE REVIEW OF THEORETICAL BACKGROUND OF CONTINUOUS IMPROVEMENT

A. Continuous Improvement Programs

Several authors revealed that CI could be achieved through practice of various methods, tools and techniques individually or in an integration approach which can facilitate the competitiveness of the organizations. The most widely applied CI programs are: Lean management [20], Six Sigma [13], Lean Six Sigma [12], [13], JIT and TQM [4], [5], [7], [11], [15], SCM [7], TPM [4], [14] to mention few of them. The existing literature demonstrates that none of the above mentioned CI initiatives is capable of solving all of the performance issues for organizations when implemented alone, however, an integrated approach become the new effective methodology in terms of attaining high-quality performance and sustainable improvements [13], [21], [22].

Moreover, the current literature provides integration of four CI programs (i.e. JIT, TQM, TPM and SCM) [9]. So that, in order to contribute new knowledge to the existing and next generation, this study integrated LSS with the other CI programs of JIT, TQM, TPM and SCM. Several authors agree that the selected CI programs have potential tools to create competitive advantage and improve global competitiveness of manufacturing industries [9], [13]. The differences and similarities between these selected CI methods is presented in Table 1 below.

TABLE I
THE DIFFERENCES AND SIMILARITIES BETWEEN CI METHODS (COMPILED BY THE RESEARCHERS FROM [23], [24])

Concepts	**JIT**	**Six-Sigma**	**Lean management**	**Lean Six Sigma**	**TQM**	**TPM**	**SCM**
Origin	Toyota Motor Corp. in 1950s	The origin is a quality evaluation by Japanese practitioners. However, developed by Motorola & dispersed by General Electric in the US	The origin is quality evaluation by Japanese practitioners and Toyota	In 1986 in the US based George group	The origin is the evaluation of quality by Japanese practitioners	M/s Nippon Denso Co. Ltd. of Japan	when the buyer-supplier understood the benefits that a cooperative relationship offers in 1980s
Theory (Aim)	Achieve zero defects, zero queues, zero inventories, zero breakdown and so on	Reducing the defects to less than 3.4 DPMO by decreasing the process variation using effective methodologies	Eliminating the waste in the process through flowing the product based on the customer demand	Eliminating waste and reducing the defects and variations in organization's processes	Focusing on satisfying the internal and external customers by armed the employees with QM tools and methodologies to achieve customer satisfaction.	Maximizing equipment effectiveness, and Achieve zero accidents, zero defects and zero breakdowns.	Improving long-term performance of the individual companies and the supply chain as a whole
Process (Concept)	Focusing on improving the flow in the process and removing all kind of waste	Focusing on reducing the process variation and improve processes	Focusing on improving the flow in the process and removing all kind of waste	Focusing on reducing the defect and process variation, and improve processes	Focusing on the improvement by organizing the process to produce the customer satisfaction.	Focusing on equipment assets management	Focusing on the entire SC starting from upstream to downstream process.
Approach	Systematic, based on planning monitoring controlling and improvement (Project management)	Systematic, based on planning monitoring controlling and improvement (Project management)	Systematic, based on planning monitoring controlling and improvement (Project management)	Systematic, based on planning monitoring controlling and improvement (Project management)	Employee's commitment with the target of the organization.	Employee's commitment with the target of the organization	Supplier's commitment with the target of the organization
Methodologies	PDCA cycle	(DMAIC Phases for process improvement) and (DMADV for developing new product	Pull system (based on the customer demand) evaluating by value stream mapping,	PDCA cycle, DMAIC or DMADV phases, VSM	(PDCA cycle) Problem-solving strategy.	PDCA cycle	PDCA cycle

Concepts	JIT	Six-Sigma	Lean management	Lean Six Sigma	TQM	TPM	SCM
		and /or process)	flow improvement and perfection				
Tools	More analytical tools integrated with quality tools	Advanced statistical and analytical tools (The advanced tools integrated with methodologies)	More analytical tools integrated with quality tools	Advanced and analytical tools integrated with methodologies and quality tools	Analytical and statistical tools.	Overall equipment effectiveness (OEE)	No tools
Primary effects	Increasing the availability of materials (raw, semi-finished or finished)	Increasing the organization bottom line and high financial orientation	Reducing the lead time in order to improve the flow by removing the waste	Reducing the lead-times and increasing the organization bottom line and high financial orientation	Increase or exceed customer satisfaction.	Increase equipment effectiveness	Increase the performance of the supply chain
Secondary effects	Achieves customer satisfaction by increasing the quality and reducing the cost of products and make the price of the products competitive	Achieving financial performance	Achieves customer satisfaction by increasing the quality and reducing the cost of products and make the price of the products competitive	Achieves customer satisfaction and financial performance	Obtains customer loyalty and improves the whole performance.	Achieve productivity, quality, cost, delivery, safety and morale	

B. Unique Practices of Integrated CI Programs

Several studies made an effort to identify unique practices for each of the aforesaid CI programs (i.e. JIT, TQM, TPM, SCM and LSS). For example, the JIT unique practices commonly identified by [4], [9], [14], [15] are setup-time reduction; JIT daily schedule adherence (time); JIT layout/Equipment layout; .JIT delivery by suppliers (frequent and reliable deliveries); pull system/Kanban. The TQM unique practices are customer focus; process management; supplier quality involvement; cross-functional product design [5], [9], [10], [14], [15], [25]. The TPM unique practices commonly identified by [9], [14] are autonomous maintenance; preventive/planned maintenance; proprietary equipment development. Other study conducted by [7] identified three SCM unique practices, namely procurement; distribution and material handling; internal logistics management. [26] also identified the six LSS unique practices specifically project management skill; project selection, prioritization, review and tracking; linking LSS to business strategy; linking LSS to the customer, linking to the suppliers; linking LSS to the process. However, these unique practices were not presented in a single framework. Thus, the study considered the above-mentioned unique practices of each CI programs in the development and validation of the proposed framework.

C. Common Success Factors for Successful Implementation of Integrated CI Framework

In the vein of unique practices, numerous studies also identified the common success factors or critical success factors. Based on the conducted extensive literature review, there are no clear success factors mentioned for successful implementation of integrated continuous improvement framework. However, apparently the common factors that were identified as human and strategic oriented common practices for JIT, TQM, TPM, SCM and LSS implementation, by many studies such as [9], [13], [14], [15] are consider as a critical success factors. These are: top management and leadership support; policy and strategy; engagement and management of people; resource management; customer focus; partnership/relationship management/linking to suppliers; utilization of problem solving approach; evidence-based decision making; information and analysis; training and education; support system integration; contemporary systems practice; organization infrastructure; effective communication; commitment for continuous improvement; review and tracking of performance; involvement of stakeholders, middle management.

In the study, these factors are categorized as social/human, strategic, technical/operational, technology, structure, resource, information and related factors based on the existing literature. Thus, the category of each critical success factors is presented as follows: Human factors (Top management & leadership support; Engagement and management of people/work force; Middle management and stakeholder involvement and commitment; Commitment for continuous improvement); Strategic factors (Policy and strategy; Partnership/relationship management; Focus on customer); Structure factors (Organization structure; Support services integration (ICT, R & D etc.)); Resource factors (Resource management; Training and education); Operational factors (Utilization of problem solving approach; Review and tracking of performance); Technology factors (Contemporary systems); Information and related factors (Evidence-based decision making; Information and analysis; Effective communication).

D. Integration Approach in Continuous Improvement

The integration approach in continuous improvement is a method of combining two or more CI methods or techniques to overcome the difficulties in a quality system and to achieve competitive advantages. [27] defined integration as a means of combining the appropriate methods and techniques to attain improvement in the operation process. [28] said that integration CI demands discipline when improving the business process to avoid the weaknesses in the CI methods. The meaning of integration in CI, according to [23], is the parallel the use of the applicable CI methods in order to achieve significant improvement in the business process while adding value to the CI system.

How methods and techniques are being integrated? Essentially, the possible approaches of integration in CI are either integrating methods with methods, techniques with techniques or methods with techniques [29]. Therefore, the literature shows that the common mechanism of the integrated method in CI is based on the possibility of the following motivations: elimination of the weaknesses in the methods or techniques, the occurrence of synergies between the homogeneity methods and or techniques and the prerequisite of enhancing one method to another in the way to exchange the results [30]. This view supported by [13], [29] who stated that elimination of the weaknesses and the possibility of occurrence of the synergies among the methods serves as the key trigger for the integrated approach.

What methods and techniques are often being integrated? According to [29] stated that the integration between CI methods and techniques can be formulated by three ways; integrating methods with methods, integrating methods with techniques and integrating techniques with techniques. According to this, the study targets on integration of methods with methods. In the current literature, the integration of CI methods with methods is not new. Different scholars made an effort to integrate two or more CI methods. For example, integration of JIT and TQM [5], [10], [11]; JIT, TQM, TPM and SCM [9]; Lean and Six Sigma [12], [13]; JIT, TQM and SCM [7]; Integration of LSS, Six Sigma and TQM [13]; JIT, TQM and TPM [4]; TPS, TQM and TPM [14]; JIT, TQM and LSS [15].

E. PDCA Cycle and DMAIC Methodology

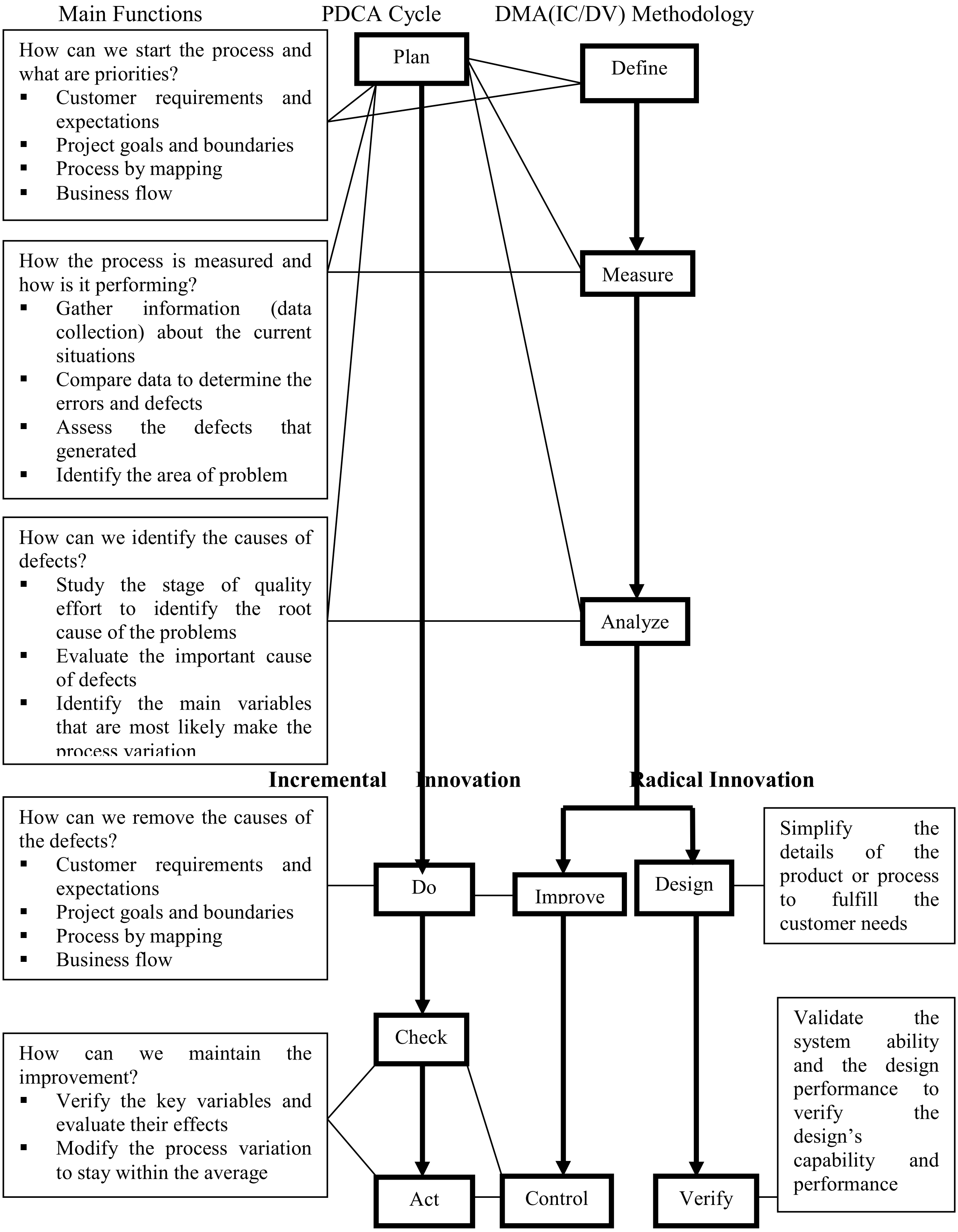
PDCA is an iterative four-step management approach used for controlling and continuously improving processes and products [14]. Reducing or eliminating organizational wastes and other problems have been approached through the Deming's or Shewhart's PDCA cycle [31]. [32] have tried to show the detailed activities of each stage of the PDCA cycle the central theme being increasing customer satisfaction. **Plan**:- Create appropriate teams, Gather all available data, Understand customers' needs, Describe the process that surrounds the Problem, Determine root cause(s), Design action plan and Develop an action plan; **Do**:-Implement improvement, Collect appropriate data, Measure progress and Document results; **Check**:- Summarize and analyze data, Evaluate results relative to targets & see Differences, Review any problems/errors, Record what was learned, Specify any remaining issues or unintended costs; **Act**:-Standardize desired improvements, Formalize current best approach, Communicate results broadly and Identify next improvement.

PDCA has got enormous applications: as a model for continuous improvement; when starting a new improvement project; when developing a new or improved design of a process; when defining a repetitive work process; and when implementing any change [14], [33]. Thus, the required functions and tools in every phases of the PDCA cycle are presented in table 2 below.

TABLE II
INTEGRATION OF PDCA CYCLE FUNCTIONS AND TOOLS (COMPILED BY THE RESEARCHERS)

PDCA Cycle	Functions	Tools & techniques
Plan	(i) Introduction of the cross functional team	Process Mapping
	(ii) Reason for selecting the theme	Bar graph, Radar chart
	(iii) Current situation analysis	Pareto diagram, Histogram
	(iv) Goal setting	Line graph
	(v) Activity plan	Gant chart
	(vi) Analysis of causes of the problem	Fish bone diagram , scatter diagram
Do	(vii) Measures examined and implemented	5W2H, Judgment criteria's
Check	(viii) Checking of results	Check sheet, Histogram, Scatter plot, Control charts
Act	(ix) Standardization and control	Control chart, check sheet
	(x) Future plan to solve another problem	5W2H

Several authors [23], [34], [35] stated that Six-Sigma comprising two main methodologies, DMAIC and DMADV in particular. DMAIC (i.e. **Define, Measure, Analysis, Improve and Control**) is the process improvement of Six-Sigma used for improving the existing process, DMADV (i.e. ***Define, Measure, Analyze, Design and Verify)*** is the other Six-Sigma methodology used for developing and design new products or processes, these methodologies are considered the roadmap of Six-Sigma deployment. Both methodologies are integrated with statistical tools and techniques. Thus, this research is also focused on integration of the first methodology with other CI methodologies. DMAIC methodology is considered to be the driving force of Six-Sigma in terms of problem-solving and sustaining the continuous improvement in particular of an existing process, DMAIC is the most popular Six-Sigma methodology based on the Deming cycle (Plan, Do, Check and Act). This cycle is used to improve existing business process [23]. The DMAIC methodology has five phases in its improvement cycle integrated with robust tools and techniques to overcome the quality problems within the system to smooth the operation's performance [36]. The function and the mechanism of these steps are described in figure (3) below.



How can we start the process and what are priorities?

- Customer requirements and expectations
- Project goals and boundaries
- Process by mapping
- Business flow

How the process is measured and how is it performing?

- Gather information (data collection) about the current situations
- Compare data to determine the errors and defects
- Assess the defects that generated
- Identify the area of problem

How can we identify the causes of defects?

- Study the stage of quality effort to identify the root cause of the problems
- Evaluate the important cause of defects
- Identify the main variables that are most likely make the process variation

How can we remove the causes of the defects?

- Customer requirements and expectations
- Project goals and boundaries
- Process by mapping
- Business flow

How can we maintain the improvement?

- Verify the key variables and evaluate their effects
- Modify the process variation to stay within the average

Simplify the details of the product or process to fulfill the customer needs

Validate the system ability and the design performance to verify the design's capability and performance

Fig. 3 Flowchart showing the relationship between PDCA and DMAIC/DV (compiled by the researchers from [23], [35], [13])

III. THE PROPOSED NEW INTEGRATED CONTINUOUS IMPROVEMENT FRAMEWORK

A. What are the Gaps Existing in the Integration of Continuous Improvement Practices

As discussed in the literature review, previously proposed frameworks (see sub section 3.4), identified unique practices (see sub section 3.2) and common factors (see sub section 3.3) for successful implementation of JIT, TQM, TPM, SCM and LSS initiatives are not corresponding. Moreover, the frameworks are at the conceptual stage and rare to find research works focus on validating the proposed frameworks. Thus, the study provides an associated result, primarily by identifying the most widely found unique and common factors, five phase implementation procedure, required tools (i.e. input box, tool box and output box), and then developing and validating the integrated framework as shown in the subsequent sections.

A framework should be simple, logical and, yet, comprehensive enough to be successful in the implementation process and attain improvement in the level of performance [13]. Accordingly, the process of developing the proposed framework is a result of integrating the unique and common factors of the five CI methods (i.e. JIT, TQM, TPM, SCM and LSS) to formulate an effective platform for conceptualization the operation system, for facilitating the five phase implementation procedures of the framework, and to provide impetus and guidance for continuous improvement and for attaining performance excellence in the manufacturing organizations. However, combination trend of PDCA cycle and DMAIC improvement methodologies are adopted as the key strategies of the framework for identifying opportunities for improvement and for obtaining the operation performance. Accordingly, the framework consists of three main components which are:

B. The main body of the framework

Similar to other studies, the structure of the framework is represented by flowchart diagram in figure (4) which displays the framework phases and steps of functions for every phases of the framework. The development of the flowchart is mainly based on identifying the literature gap and conducted a literature review to fill the gap as described in the above sections.

C. The main elements of the framework

The unique and common factors (a described in the above sections), input box, toolbox and output box (see Appendix A) containing set of tools and techniques to formulate the five phases of the framework, and organization competitiveness measures are the main elements encompassed in the framework. They are also prioritized and organized based on combination trend of PDCA and DMAIC methodologies to deliver the tasks phase by phase in order to provide the opportunity for operational, innovation and business improvement and to overcome competitiveness problems. This is supported by [13] as he identified and considered different lean, six sigma and TQM tools and techniques in every five phases of the integrated QM framework. The development of these elements is based on the literature review in the above sections.

D. The operational mechanism of the framework

Every steps of each phases, unique and common practices, work activities based on the identified steps of functions of the framework are organized based on the trend of combination of PDCA cycle and DMAIC methodology, where the phases of the framework are integrated to gather and to simplify the operation process and to enhance competitiveness. This mechanism is designed as an integrated and unified system to operate the framework, the development of these integrated functions and mechanism is based on the literature review in above sections.

This new framework is quite different from the existing JIT and TQM; JIT, TQM and TPM; JIT, TQM and SCM; Lean and Six Sigma; Six-Sigma and TQM; combination of LSS and Six Sigma-TQM; JIT, TQM and LSS; JIT, TQM, TPM and SCM integrated frameworks recommended by existing literatures (e.g., see the frameworks suggested by [4], [5], [7], [10], [11], [13], [14], [25].

Thus, the new framework encompassed ten additional features to the existing frameworks as illustrated in figure 4. These additional features of the new framework are presented as follows: The integrated framework is based on the trend of combination of PDCA approach and DMAIC methodology (i.e. the integrated framework is developed in considering the work activities, the functions and tools and techniques, social and technical factors of the PDCA cycle and DMAIC methodology). It considers structured implementation procedure with five different phases. It considers the requirements of twenty one unique practices of the five continuous improvement initiatives including LSS, JIT, TQM, SCM and TPM. It considers fundamental seventeen common success factors including support systems integration and contemporary system practices. It considers comprehensive tooling of the integrated framework by assigning different tools and techniques as input box, toolbox and output box in every phase of the integrated framework (see Appendix A). It considers three key players (i.e. manufacturing industries, institutions, and government) with their roles for successful implementation of the integrated framework and completion of the transformation. It considers evidence-based decision making as one additional common principle of the integrated framework. It also provides consolidated operational (customer, quality, productivity (including cost, delivery and flexibility), people, society, maintenance, supplier and resource results); innovation (new product innovation, improved existing product, new process innovation, improved existing process, new market innovation, new organization's issues); and business results (marketing relating to size of sales and market share, financial relating to level of profit, company's earnings, return on investment and inventory turnover) as key competitiveness indicators (KCIs) to measure the impact of the proposed framework. It considers initial expert team building before starting execution of a pilot project. It considers policy and strategy development and deployment to respective sections or departments based on conducting the situational analysis using SWOT analysis and other instruments.

These ten additional features of practices are very useful for manufacturing firms operating in developing countries (For example Ethiopia). They are also missed in the existing literature due perhaps to the fact that the existing literatures are more leaned towards the situations of the firms of developed nations. The proposed framework is an innovative framework as it helps manufacturing industries to base their organizational performance, business success and competitiveness on integrated continuous improvement practices of JIT, TQM, TPM, SCM and LSS knowledge transfer from external sources such as institutions (i.e. Ethiopian Kaizen institute, universities, ministry of trade and industry developmental institutes and others) as well as internal sources such as existing organizational performance measures based on assessment and review.

Moreover, manufacturing industries require successful integrated CI framework to enhance the strengths of each CI methods, which in turn it successively brings about success and competitiveness through continual improvement effort. Thus, the proposed framework is ground-breaking framework as it helps manufacturing industries to secure their profit and will add value to both academicians and practitioners who are consulting on the manufacturing industries. In general, due to its superior features, this new framework can create better insight into the scientific audience and industrial applications. Thus, the proposed framework is based on the full implementation of five phases (see figure 4) and it overcomes the limitation of the existing frameworks under the implementation framework category of integrated continuous improvement framework.

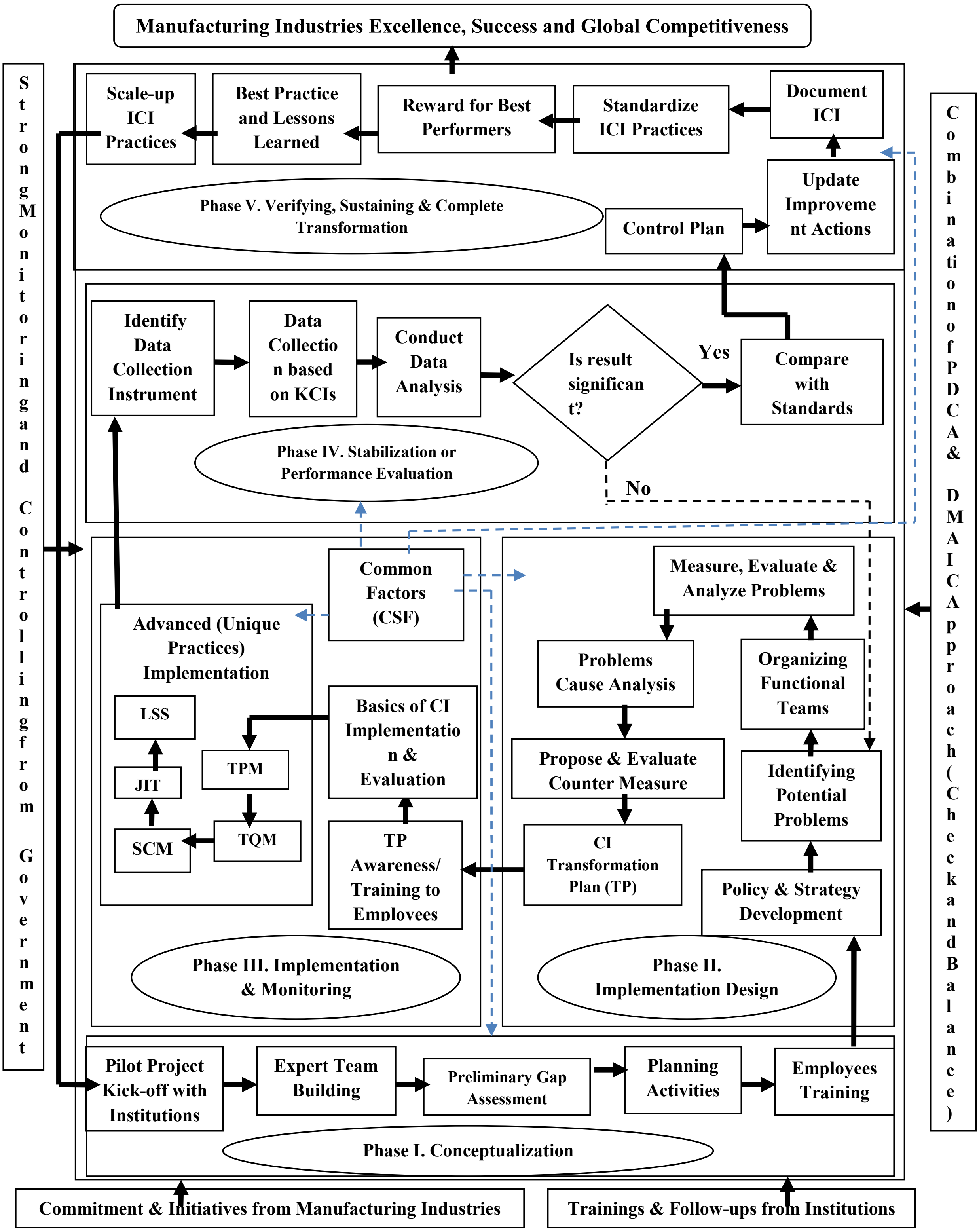

Manufacturing Industries Excellence, Success and Global Competitiveness
Scale-up ICI Practices
Best Practice and Lessons Learned
Reward for Best Performers
Standardize ICI Practices
Document ICI
Update Improveme nt Actions
Control Plan
Phase V. Verifying, Sustaining & Complete Transformation
Identify Data Collection Instrument
Data Collectio n based on KCIs
Conduct Data Analysis
Is result significan t?
Yes
No
Compare with Standards
Phase IV. Stabilization or Performance Evaluation
Common Factors (CSF)
Advanced (Unique Practices) Implementation
LSS
JIT
SCM
TPM
TQM
Basics of CI Implementatio n & Evaluation
TP Awareness/ Training to Emplovees
Phase III. Implementation & Monitoring
Measure, Evaluate & Analyze Problems
Problems Cause Analysis
Organizing Functional Teams
Propose & Evaluate Counter Measure
Identifying Potential Problems
CI Transformation Plan (TP)
Policy & Strategy Development
Phase II. Implementation Design
Pilot Project Kick-off with Institutions
Expert Team Building
Preliminary Gap Assessment
Planning Activities
Employees Training
Phase I. Conceptualization
Strong Monitoring and Controlling from Government
Combination of PDCA & DMAIC Approach (Check and Balance)
Commitment & Initiatives from Manufacturing Industries
Trainings & Follow-ups from Institutions

Fig. 4 A new project-based integrated continuous improvement framework (the researchers)

IV. VALIDATION OF THE PROPOSED INTEGRATED CONTINUOUS IMPROVEMENT FRAMEWORK

This section contains a statistical and graphical analysis of the data collected from the sample manufacturing companies. The findings and detail discussion are presented according to the research objectives and research questions given in the introduction part of this study. The main focus of this chapter is to discuss the questionnaire survey that has been developed to verify and validate the proposed framework that was presented in the previous section along with how the data collected and analyzed. This enables the researcher to modify the proposed framework if necessary and can also enhance and increase the confidence level of the researcher with respect to developing the framework. The results of the questionnaires are provided for each section and the validation steps were performed using SPSS to confirm the validity and reliability of the framework and its procedures for implementation.

A total of 350 questionnaires were sent to 50 companies, which are located in the Addis Ababa and rounds. Seven questionnaires were sent to each organization and the respondents were the senior level managers from different departments in the companies. A total of 42 questionnaires were not filled by the respondents from 9 companies and 308 completed questionnaires were received from 41 companies. Thus, the response rate was 82.00% company wise and 88.00% respondent wise. The analysis of the data was conducted in August 2021. As discussed in section two, the perception based data is collected for this study. Multiple responses are collected from a single organization. However, as argued by [36] that different informants from the same organization might have different opinion on the same issue due to the difference in their perceptions and knowledge therefore each respondent is considered as an independent case for the further analysis. The evaluation of the adequacy of the data collection instrument is the fundamental aspect of any analysis; thus, before starting the data analyses the reliability and validity of the research instrument are described. Subsequently, the profile of the respondents and sampled companies is reviewed.

A. Integrity Data Analysis (Reliability and Validity) of the Questionnaire

In this section, the reliability and validity of the questionnaire are evaluated, based on the data collected. The adequate reliability and validity analysis provides the confidence to the audience that the findings are reliable and are based on the accurate measures of the underlying constructs.

a. Reliability Analysis of the Proposed Framework

Several authors highlighted that checking the reliability of data's internal consistency before proceeding to descriptive analysis, factorial analysis [37], [38]. A Reliability test is a crucial measure to assess the quality of the instruments used in the questionnaire and to check the reliability of data collected, Cronbach Alpha was also undertaken in this section to measure the internal consistency of the instruments used to evaluate the proposed model. Ideally, Cronbach alpha must be greater than 0.7 to consider the items being measured are consistent and reliable [13], [39]. Therefore, the test was carried out for each of the eight statements used to evaluate the proposed framework. Based on this, the results in table (3) demonstrated that the coefficient alpha and the standardized item alpha for eight statements are 0.741 and 0.713 respectively which all are greater than 0.70, accordingly that is an indication that all of the items are consistent and reliable.

TABLE III
RELIABILITY STATISTICS

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.741	.713	8

However, the results in table (4), column three labeled '*corrected item-total correlation*' showed that there is positive correlation between the whole items except item number eight 'Evaluating the proposed framework in terms of anything missing and should be added to the proposed framework' which has negative correlation with value (-0.062). In addition, in column five labeled 'Cronbach's alpha if item deleted' the same item has the highest alpha value, 0.917. Accordingly, if item number eight were deleted from the calculation, then Cronbach alpha would be improved.

TABLE IV
ITEM-TOTAL STATISTICS

	Scale Mean if Item Deleted	*Scale Variance if Item Deleted*	*Corrected Item-Total Correlation*	*Squared Multiple Correlation*	*Cronbach's Alpha if Item Deleted*
Appropriateness & applicability (F1)	*19.63*	*12.873*	*.807*	*.629*	*.723*
Ability to boost competitiveness (F2)	*19.64*	*13.971*	*.738*	*.615*	*.710*
Ability to overcome problems (F3)	*19.81*	*13.897*	*.693*	*.532*	*.702*
Ability to overcome complex nature of CI implementation (F4)	*19.83*	*12.572*	*.716*	*.697*	*.718*
Ability to achieve long-term goals (F5)	*19.89*	*13.027*	*.727*	*.583*	*.737*
Role of three key players (F6)	*19.69*	*13.897*	*.771*	*.471*	*.726*
Combination of PDCA & DMAIC (F7)	*19.91*	*12.481*	*.683*	*.492*	*.730*
Anything missed & should be added (F8)	*21.93*	*17.295*	*-.062*	*.089*	*.917*

After deleting item number eight and running the test again, the results in table (5) below indicated that Cronbach alpha is 0.793 and that the standardized item alpha is 0.794.

TABLE V
RELIABILITY STATISTICS

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.793	.794	7

Additionally, in table (6), in Column three all the items are correlated with value above 0.3 and in column five value of Cronbach alpha if items deleted ranged between (0.76 to 0.78) which is greater than 0.7 Subsequently it can be conclude that the entire instruments have high internal consistency and reliable.

TABLE VI
ITEM-TOTAL STATISTICS

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
F1	17.28	12.311	.801	.624	.781
F2	17.09	13.173	.733	.610	.774
F3	17.31	16.730	.679	.527	.785
F4	18.27	12.730	.711	.692	.763
F5	17.53	12.592	.723	.578	.761
F6	17.06	12.853	.769	.466	.773
F7	17.28	12.574	.673	.487	.780

b. Validity Test and Validation the Proposed Framework

Validity tests confirm the degree to which the measures used in the study are truthfully measuring what is intended to be measured [40]. As they should be performed to check the accuracy and truthfulness of the results, Chi-square goodness of fit (x^2) was applied to check the validity of the instruments that were used to evaluate the proposed framework. Chi-square goodness of fit is used to find out whether an observed value is statistically, significantly different from the expected value (Field, 2009). The Chi-square goodness of fit with corresponding P value is considered to be significant if P value ≤ 0.05 [41].

As can be seen in table (7), the results of x^2 demonstrated that the P values are less than 0.05, which means that the results are significantly different from the actual observed values and the expected values of all the statements used to evaluate the proposed model. That also can be an indication for the possibility of publishing the results and generalizing from the current research sample to the entire publication [13].

TABLE VII
TEST STATISTICS

	F1	F2	F3	F4	F5	F6	F7	F8
Chi-Square	65.032[a]	55.260[a]	57.039[a]	55.337[a]	56.852[a]	55.931[a]	58. 573[a]	57.109[b]
df	3	3	3	3	3	3	3	1
Asymp. Sig.	.000	.000	.000	.000	.000	0.000	0.000	.000

a. 0 cells (0.0%) have expected frequencies less than 5. The minimum expected cell frequency is 13.3.
b. 0 cells (0.0%) have expected frequencies less than 5. The minimum expected cell frequency is 55.0.

B. Descriptive Analysis

This section provides the descriptive analysis of the data collected using SPSS 23. Various descriptive measures were used to measure the central tendency (mean), allowing the results of data analysis to be provided in the following sections in forms of tables, charts and different statistics and figures.

a. Background Information Analysis

The section provides the results of the questionnaires received from the respondents. The aim of this part of the survey is to present a clear picture of the respondent's background and to understand the awareness level of the existing continuous improvement in the organization.

Respondent's position: The respondents were asked to state their position within their organization. The results listed in the table (8) showed that 15.6% of the respondents are production managers, 14.9% are quality assurance managers, 14.0% are HR managers. Besides, 14.3%, 14.3%, 14.0%, 13.0% are sales & marketing managers, supply & procurement managers, product development managers and finance managers respectively.

TABLE VIII
POSITION WITHIN THE ORGANIZATION

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Production manager	48	15.6	15.6	15.6
	Quality assurance manager	46	14.9	14.9	30.5
	HR manager	43	14.0	14.0	44.5
	Sales & Marketing manager	44	14.3	14.3	58.8
	Supply & Procurement manager	44	14.3	14.3	73.1
	Product development center manager	43	14.0	14.0	87.0
	Finance manager	40	13.0	13.0	100.0
	Total	308	100.0	100.0	

Area of industry: The respondents were asked to indicate the manufacturing sub sector in which their organizations functioned; the results were shown in a table (9). 18.2% of the respondents belong to the agro processing, 13.6% belong to the metal engineering, 11.4% belong to the textile and 11.4% are from garment industry. The other manufacturing subsectors of leather, leather products, automotives, chemical and pharmaceutical scored 11.0%, 11.0%, 8.1%, 11.0% and 4.2% respectively.

TABLE IX
AREA OF INDUSTRY

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Textile	35	11.4	11.4	11.4

	Garment	35	11.4	11.4	22.7
	Leather	34	11.0	11.0	33.8
	Leather product	34	11.0	11.0	44.8
	Metal engineering	42	13.6	13.6	58.4
	Automotive	25	8.1	8.1	66.6
	Agro-processing	56	18.2	18.2	84.7
	Chemical	34	11.0	11.0	95.8
	Pharmaceutical	13	4.2	4.2	100.0
	Total	308	100.0	100.0	

The type of the continuous improvement initiative currently employed in respondent's organizations: The respondents were asked to indicate the continuous improvement program currently used within the organization, the results in table (10) demonstrated as follows; 47.7% TQM, 18.2% SCM, 8.4% TPM, 1.3% JIT, 0% LSS and 24.4% other continuous improvement initiatives.

TABLE X
THE CURRENT CONTINUOUS IMPROVEMENT INITIATIVE OF THE RESPONDENT'S ORGANIZATION

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	JIT	4	1.3	1.3	1.3
	TQM	147	47.7	47.7	49.0
	TPM	26	8.4	8.4	57.4
	SCM	56	18.2	18.2	75.6
	LSS	0	0	0	75.6
	Others	75	24.4	24.4	100.0
	Total	308	100.0	100.0	

The level of awareness with the following continuous improvement initiative tools/techniques: Various tools and techniques relating to JIT, TQM, TPM, SCM and Lean Six Sigma initiatives were presented to the respondents and they were asked to indicate if they were aware of any of the tools/techniques listed in the survey questions. The results presented in table (11) showed that the majority of respondents seem to be familiar with most of tools and indicated that the level of the awareness with these tools were above 50% which are slightly above average. However, the only tools ranked below average are 'Taguchi methods', 'Force field study', 'Process capability analysis', 'PERT chart', 'DOE', 'ANOVA', 'QFD', 'DMAIC Process', 'VSM', 'SIPOC analysis', 'FMEA' and 'Regression analysis'.

TABLE XI
THE AWARENESS LEVEL WITH JIT, TQM, TPM, SCM and LSS TOOLS/TECHNIQUES IN THE MANUFACTURING INDUSTRIES

Tools/Techniques name	**Frequency**		**Percentage (%)**	
	Yes	***No***	***Yes***	***No***
Project charter	*166*	*142*	*53.9*	*46.1*
Customer surveys	*298*	*10*	*96.8*	*3.2*
Quality function deployment (QFD)	*40*	*268*	*13.0*	*87.0*
PDCA (plan, do, check, act)	*206*	*102*	*66.9*	*33.1*
DMAIC Process (define-measure-analysis-improve-control)	*35*	*273*	*11.5*	*88.5*
Process flow chart/mapping	*30*	*278*	*9.6*	*90.4*
Statistical Process Control (SPC)	*193*	*115*	*62.7*	*37.3*
Process capability analysis	*3*	*305*	*1.0*	*99.0*
Quality control circles (QCCs)	*308*	*0*	*100.0*	*0.0*
Pareto analysis	*192*	*116*	*62.3*	*37.7*

Tools/Techniques name	Frequency		Percentage (%)	
	Yes	*No*	*Yes*	*No*
Histogram	*165*	*143*	*53.6*	*46.4*
Design of experiments (DOE)	*5*	*303*	*1.6*	*98.4*
Benchmarking	*168*	*140*	*54.5*	*45.5*
SIPOC (Suppliers, Input, Process, Output, Customers)	*57*	*251*	*18.5*	*81.5*
Taguchi methods	*0*	*308*	*0.0*	*100.0*
Single minute exchange of die (SMED)	*177*	*131*	*57.5*	*42.5*
Kanban/Line balancing	*192*	*116*	*62.3*	*37.7*
Suggestion box	*308*	*0*	*100.0*	*0.0*
Value stream mapping (VSM)	*48*	*260*	*15.6*	*84.4*
Regression analysis	*97*	*211*	*31.5*	*68.5*
Brainstorming techniques	*302*	*6*	*98.1*	*1.9*
Cause and effect diagram/analysis	*269*	*39*	*87.3*	*12.7*
Total productive maintenance (TPM)	*166*	*142*	*53.9*	*46.1*
Poka-yoke	*159*	*149*	*51.6*	*48.4*
Team building methods	*261*	*47*	*84.7*	*15.3*
Root causes analysis	*257*	*51*	*83.4*	*16.6*
Run charts	*229*	*79*	*74.3*	*25.7*
Kaizen	*308*	*0*	*100.0*	*0.0*
PERT chart (program evaluation and review technique)	*6*	*302*	*2.0*	*98.0*
Force field analysis	*0*	*308*	*0.0*	*100.0*
Failure mode and effect analysis (FMEA)	*63*	*245*	*20.5*	*79.5*
Analysis of variation (ANOVA)	*4*	*304*	*1.3*	*98.7*
Quality cost systems - cost of poor quality (COPQ)	*168*	*140*	*54.5*	*45.5*
Supplier audit	*283*	*25*	*91.8*	*8.2*
5S	*308*	*0*	*100.0*	*0.0*

b. Validation of the Proposed Framework

This section of the survey seeks to validate the proposed framework for manufacturing industries, it aims to provide an understanding of the implementation procedures suitable for manufacturing organizations, identify the difficulties in implementing the proposed framework and reveal the accuracy level in its contents in terms of helping manufacturing organizations to gain a competitive advantage in the long run. The framework was presented to the respondents and were asked to evaluate the framework in terms of the suitability for manufacturing organization and their applicability to achieve competitive advantages, additionally, evaluation the implementation procedures of the framework for manufacturing organizations was based on the ranking below; *1-Strongly Disagree (SD). 2-Disagree (D). 3-Moderate (M). 4-Agree (A). 5-Strongly Agree (SA).* The results of the evaluation of the proposed framework were as follows. The results in table (12) demonstrated that F1, F2, F3, F4, F5, F6, F7 and F8 are supported by 70.1%, 77.9%, 83.5%, 77.9%, 73.4%, 89.9%, 77.96% and 89.94% of the respondents.

TABLE XII
THE EVALUATION OF THE PROPOSED FRAMEWORK

Questions regarding to proposed framework	Results of respondents in %				
	SD	D	M	A	SA
Appropriateness and applicability (F1)	1.9	3.6	24.4	64.6	5.5
Ability to boost excellence & competitiveness (F2)	1.3	2.9	17.9	69.2	8.7
Ability to deal with and to overcome problems (F3)	3.2	2.3	11.0	77.0	6.5
Ability to overcome complex nature of ICI implementation (F4)	2.6	3.9	15.6	66.2	11.7
Ability to achieve long-term goals & business expectations (F5)	4.2	2.9	19.5	61.4	12.0
Role of manufacturing industries, institutions & government (F6)	1.6	5.2	3.3	70.1	19.8

Overall combination of PDCA and DMAIC methodology (F7)	2.9	5.5	13.6	61.4	16.6
Anything missing & should be added to the framework (F8)	Yes (10.06) and No (89.94)				

c. Validation the Implementation Procedures of the Framework

This part of the research seeks to evaluate the procedures concerning implementation of the proposed framework with respect to every step encompassed in five phases see section (4). The procedures designed for implementation were presented to the respondents in the framework. The respondents were asked to indicate how the statements related to each phase based on the following ranking: *1-Strongly Disagree (SD). 2-Disagree (D). 3-Moderate (M). 4-Agree (A). 5-Strongly Agree (SA).*

The results in table (13) demonstrated that 84.4%, 81.1%, 83.4%, 89.6%, 83.4%, 90.2% and 92% of the respondents were supported and in agreement with the contents of the phase one, phase two, phase three, unique practices, common factors, phase four and phase five respectively. In addition, the table (13) signified that 79.5%, 79.2%, 90.5%, 78.3%, 90.3%, 94.5%, 88.0% of the respondents was supported and in agreement with the applicability of the five functions in phases one, seven functions in phase two, four functions in phase three, steps of unique practices, common factors, four functions in phase four and seven functions in phase five respectively in assisting and success the conceptualization, implementation design, implementation and monitoring, performance evaluation , and verifying and sustaining process of the implementation procedure. All most all the respondents agreed that there are no elements missing from the phase one conceptualization to phase five of verifying, sustaining and complete CI transformation.

TABLE XIII

THE EVALUATION OF THE PROPOSED IMPLEMENTATION PROCEDURE

Questions regarding to implementation procedure	Results of respondents in %				
	SD	D	M	A	SA
Evaluating the contents (five functions) of phase one			15.6	75.7	8.7
Evaluation the applicability of five functions in phase one			20.5	66.9	12.6
Evaluation the phase one in terms of anything missed	Yes (0.0), No (100)				
Evaluation the contents (seven functions) of phase two	1.3	2.0	15.6	75.3	5.8
Evaluation the ability of phase two to achieve the key target	2.6	1.6	16.6	68.2	11.0
Evaluation the phase two in terms of anything missed	Yes (3.6), No (96.4)				
Evaluation of the contents (four functions) of phase three	3.9	1.6	11.0	72.7	10.7
Evaluation the ability of phase three to achieve key target	1.9	4.9	2.6	37.3	53.2
Evaluation the phase thee in terms of anything missed	Yes (2.6), No (97.4)				
Evaluation the contents of unique practices of CI initiatives			10.39	71.7	17.9
Evaluation the prerequisite practice of unique practices			21.7	69.2	9.1
Evaluation the unique practices in terms of anything missed	Yes (0.0), No (100)				
Evaluation of the contents of common success factors			16.6	51.30	32.14
Evaluation ability of common factors to achieve key target			9.7	53.9	36.4
Evaluation the common factors in terms of anything missed	Yes (0.0), No (100)				
Evaluation the contents (four functions) of phase four		1.0	8.8	78.6	11.7
Evaluation the ability of phase four to achieve the key target			5.5	77.6	16.9
Evaluation the phase four in terms of anything missed	Yes (0), No (100)				
Evaluation of the contents (seven functions) of phase five	0.6	2.3	5.1	80.0	12.0
Evaluation the ability of phase five to achieve the key target		1.6	10.3	81.5	6.5
Evaluation the phase five in terms of anything missed	Yes (1.6), No (98.4)				

C. Discussion on the Findings

The research investigated two main issues; one is evaluating and validating the proposed framework and its implementation procedures. The validation process is undertaken using the quantitative approach represented by the questionnaire survey. It was carried out after the development of the framework; the questionnaire survey involved 308 managers in the industrial sector. It sought to obtain the opinions and ideas of managers in terms of the suitability of the framework for manufacturing organizations. The framework was assessed in terms of; the appropriateness and applicability, the capability and effectiveness, the role of the three key players (manufacturing industries, institutions, and government) to successful implementation of the framework and the overall combination of PDCA and DMAIC methodology to check and balance during the implementation of the framework.

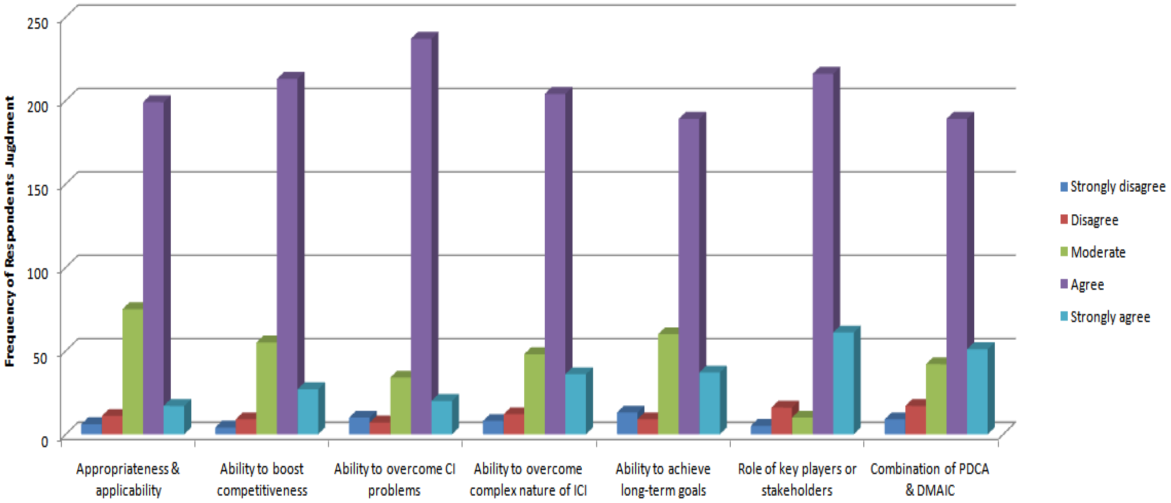


Fig. 5 Respondents Judgment about the evaluation of the framework

The results confirmed that the proposed framework developed is applicable for manufacturing organizations and can assist in achieving competitive advantages if adopted or applied correctly, figure 4 provided evidence from the research outcomes in which it is clearly demonstrated that a very high percentage of respondents agreement with contents, appropriateness, competitive advantages, effectiveness and completeness of the framework. Thus, considerable attention should be paid if the framework is implemented.

V. CONCLUSION

In the last few decades the concepts and frameworks of continuous improvement were warmly welcomed and most recently, there are few integrated frameworks developed to improve the performance and secure global competitiveness of manufacturing industries. To mention few, integration of JIT, TQM, TPM and SCM; integration of Lean Six Sigma model and Six Sigma - TQM model; JIT, TQM and SCM; JIT, TQM and TPM; JIT, TQM and LSS. However, there is no study proposing and validating the integrated CI implementation framework of JIT, TQM, TPM, SCM and LSS practices with the context of manufacturing industries in Ethiopia and other countries.

To fill this gap, this study aims to develop and validate an integrated CI implementation framework for manufacturing industries in order to eradicate the competitiveness serious issues and make the integrated CI system in position more effectual for the industries. To achieve this purpose and to propose and validate the framework, the study performed three main tasks. These are: extensive reviewed of literature, proposed and validated the integrated implementation framework.

Thus, in response to RQ1 and to mitigate this problem, this study employed exploratory qualitative approach and reviewed the most widely cited unique and common practices JIT, TQM, TPM, SCM & LSS initiatives and frameworks, as a result, the study developed new integrated CI implementation framework with five phases in considering combination trend of PDCA cycle and DMAIC methodology: conceptualization; implementation design including strategic planning; implementation & monitoring; performance evaluation or stabilization; verifying, sustaining & complete CI transformation with thirty one steps under the five phases as depicted in Figure (4). The unique practices and common factors are the two functions considered in the integrated framework as unique practices in phase three and common factors in the entire phases. It includes five JIT unique practices, four TQM, three TPM, three SCM, six LSS unique practices and seventeen common practices categorized under seven factors as described in the literature review sections. Moreover, this study identified input box, tool box and output box (see Appendix A) as per the nature of the five phases and their required steps to complete each phases.

Furthermore, in response to RQ2 and to mitigate this problem, this study used quantitative approach represented by the questionnaire survey. It was carried out after the development of the framework; 308 completed questionnaires were received from 41 companies with response rate of 82.00%. The results confirmed that the proposed framework is applicable for manufacturing industries and can assist in achieving competitive advantages when the manufacturing industries offer unconditional efforts in employing the full contents of the framework and collaboration among three key players (i.e. manufacturing industries, institutions and government) play their roles as depicted in the proposed framework. Hence, manufacturing industries have to undergo a total revolution in all the dimensions of the integrated JIT, TQM, TPM, SCM & LSS practices proposed in this research in order to improve their performance, business success and global competitiveness.

VI. IMPLICATIONS OF THE RESEARCH

This is the only comprehensive study about developing the joint implementation of JIT, TQM, TPM, SCM and LSS practices as a single Integrated Continuous Improvement Framework (ICIF) with the category of implementation framework in the context of Ethiopian manufacturing industries. Thus, this study contributes into the body of ICI knowledge by providing new integrated continuous improvement implementation framework. It was identified in section 3, 4 and 5 that many gaps exist in the literature on the development of the framework in the context of the manufacturing industries. This study has tried to fill these gaps. Thus, the most important contributions of this research to the existing body of knowledge of CI are provided in relation to theoretical and managerial implications as follows.

A. Theoretical Implications

In relation to theoretical implication, this study provided the first ever an integrated continuous improvement framework of JIT, TQM, TPM, SCM and LSS practices build in five phases with different steps/functions in every phase to ascertain a distinctive continuous improvement strategy in implementation with the objective of attaining competitiveness in operation, innovation and business performance within the manufacturing industries. This study also provided an implementation procedure with thirty one functions or steps (i.e. five in conceptualization; seven in implementation design; eight in implementation & monitoring; four in performance evaluation or stabilization; seven in verifying, sustaining and complete CI transformation phase. This study provided twenty one unique and seventeen common practices or critical success factors for successful implementation of the proposed integrated continuous improvement framework with the category of implementation framework. This study provided fourteen organization competitiveness/performance indicators for measuring the effect of the proposed integrated continuous improvement implementation framework.

B. Managerial Implications

In relation to managerial implications, this study contributes to a better understanding on the potential effects of unique and common practices of JIT, TQM, TPM, SCM & LSS initiatives can have in improving operational, innovation and business performance; hence, the developed implementation framework may serve as a guideline for the manufacturing industries managers, institutions and government. Based on the results of this research, some suggestions are made for manufacturing industry managers.

Managers should give weight to the proposed framework including different unique and common practices of JIT, TQM, TPM, SCM & LSS initiatives to have an effective integrated CI implementation framework and improved operation, innovation and business performance.

VII. LIMITATIONS AND FUTURE RESEARCH DIRECTIONS

The limitations of the present study provide directions for future research as follows. Since the study is limited to small number of manufacturing industries, so future studies could survey same sector (leather and leather products, textile and garment, chemical, pharmaceutical, metal engineering, cement, sugar, agro-processing) in Ethiopia and others through self-administered questionnaire with large population data to test the practicability of the implementation framework. The study did not collect qualitative data from the required several manufacturing organizations, academicians and consultants by employing interview and focus group discussion. Thus, further qualitative investigation is required on the framework involving several manufacturing organizations, academicians and consultants and an adequate sample size to check the impact of the framework on the CI system within manufacturing industries. This could enhance and improve the contents and structure of the framework. The study did not conduct application of the integrated framework in manufacturing industries. Thus, further study can be carried out in manufacturing industries by applying the proposed ICI framework based on the designed implementation procedure in order to test its effectiveness on achieving performance results.

REFERENCES

[1]. Bhuiyan, N., Baghel, A., “An Overview of Continuous Improvement: From the Past to the Present,” Management Decision, vol. 43, Issue 5, pp. 761-771, 2005.

[2]. Singh, J., Singh, H., “Kaizen philosophy: a review of literature,” The IUP Journal of Operations Management, vol. 8, Issue 2, pp. 51-72, 2009.

[3]. Hagos, A.M., Kahsay, G., “Implementation of Continuous Improvement (Kaizen) Tools and Its Challenges in Garment Factories: Case of MAA Garment and Textile Factory (Master’s thesis),” University of Mekelle, Ethiopia Institute of Technology, 2011.

[4]. Cua, K. O., McKone, K. E., Schroeder, R. G. “Relationships between implementation of TQM, JIT, and TPM and manufacturing performance,” Journal of operations management, vol. 19, Issue 6, pp.675-694, 2001.

[5]. Tesfaye G., Kitaw, D., “A TQM and JIT Integrated Continuous Improvement Model for Organizational Success: An Innovative Framework,” Journal of Optimization in Industrial Engineering, vol. 10, Issue 22, pp. 15-23, 2017.

[6]. Sahoo, S., Yadav, S., “Influences of TPM and TQM Practices on Performance of Engineering Product and Component Manufacturers,” Journal of Procedia Manufacturing, vol. 43, pp. 728-735, 2020.

[7]. Dametew, A. W., Kitaw, D., Ebinger, F., “Enhancing Basic Metal Industry Global Competitiveness through Total Quality Management, Supply Chain Management and Just-In-Time,” Journal of Optimization in Industrial Engineering, vol. 13, Issue 2, pp. 27-46, 2020.

[8]. Snee, R.D., “Lean Six Sigma-getting better all the time,” International Journal of Lean Six Sigma, vol. 1 Issue 1, pp. 9-29, 2010.

[9]. Nandurkar, K.N., Wakchaure, V.D., Kallurkar, S.P., “The Simulation-Based Comparison of Joint Implementation of JIT, TQM, TPM and SCM Methods,” Proceedings of NAMRI/SME, pp. 42, 2014.

[10]. Getachew, F., Mulugeta, A., Kitaw, D., Berhan, E., “A synergistic frame-work of JIT & TQM through TOC,” Journal of Multidisciplinary Engineering Science and Technology, vol. 4, Issue 4, pp. 6995-7003,

2017.

[11]. Dametew, A. W., Kitaw, D., Ebinger, O., "The Roles of TQM & JIT for basic metal industries global competitiveness," Industrial Engineering and Management, vol. 6, Issue 2, pp. 1-12, 2017.

[12]. Timans, W., "Continuous quality improvement based on Lean Six Sigma in manufacturing small and medium sized enterprises," University of Groningen, SOM research school, 2014.

[13]. Khamkham, M., "Development of an Integrated Quality Management Framework for Manufacturing Organizations," Sheffield University, Doctoral Thesis, 2017.

[14]. Hailu, H., Mengstu, S., Hailu, T., "An Integrated Continuous Improvement Model of TPM, TPS and TQM for Boosting Profitability of Manufacturing Industries: An Innovative Model & Guideline," Management Science Letters, vol. 8, pp. 33-50, 2018.

[15]. Sisay, G., Kitaw, D., Ebinger, F., Jilcha, K., "Developing Integrated Continuous Improvement Model for Competitiveness of Ethiopian Automotive Industry," European Online Journal of Natural and Social Sciences, vol. 10, Issue 2, pp. 223-247, 2021.

[16]. Taddese, F., "Application of TQM for innovation: An exploratory research of Japanese, Indian and Thailand companies," International Journal of Innovation and Technology Management, vol. 14, Issue 4, pp. 1-20, 2017.

[17]. Hailu, H., Taddese, F., Tsegay, K., Jilcha, K., Hailu, T., "Relationship between Kaizen Philosophy and Organizational Performance Empirical Investigation: a Case of Ethiopian Manufacturing Industries," European Online Journal of Natural and Social Sciences, vol. 9, Issue 4, pp. 735-751, 2020.

[18]. Berhe, H.H., "Application of Kaizen Philosophy for Enhancing Manufacturing Industries Performance: Exploratory Study of Ethiopian Chemical Industries," International Journal of Quality and Reliability Management, vol. 39, Issue 1, pp. 204-235, 2022. doi: 10.1108/IJQRM-09-2020-0328

[19]. Saunders, M., Lewis, P., Thornhill, A., Wang, C., "Analyzing qualitative data, Research methods for business students," 5th edition. Harlow, Essex, UK: Pearson Education Ltd, pp. 480-525, 2009.

[20]. Lizarelli, F.L., Toledo, J.C.d., Alliprandini, D.H., "Relationship between continuous improvement and innovation performance: an empirical study in Brazilian manufacturing companies," Total Quality Management and Business Excellence, pp. 1-24, 2019.

[21]. Antony, J., "Six sigma vs. TQM: Some perspectives from leading practitioners and academics," International Journal of Productivity and Performance Management, vol. 58, Issue 3, pp. 274-279, 2009.

[22]. Johannsen, F. "A holistic approach for integrating methods in quality management," In Wirtschaftsinformatik, pp.63, 2013.

[23]. Andersson, R., Eriksson, H., Torstensson, H., "Similarities and differences between TQM, six sigma and lean," The TQM Magazine, vol.18, Issue 3, pp. 282-296, 2006.

[24]. Ahuja I. P. S., Khamba J. S. "An evaluation of TPM initiatives in Indian industry for enhanced manufacturing performance," International Journal of Quality and Reliability Management, vol. 25, Issue 2, pp. 147-172, 2008.

[25]. Flynn, B.B., Sakakibara S., Schroeder, R.G., "Relationship between JIT and TQM: practices and Performance" Academy of management journal, vol. 48, pp. 1325-1360, 1995.

[26]. Laureani, A., Antony, J., "Critical success factors for the effective implementation of Lean Sigma: results from an empirical study

and agenda for future research," International Journal of Lean Six Sigma, vol. 3, Issue 4, pp. 274-283, 2012.

[27]. Gijo, E., Rao, T. S., "Six sigma implementation-hurdles and more hurdles," Total Quality Management and Business Excellence, vol. 16, Issue 6, pp. 721-725, 2005.

[28]. Bendell, T., "A review and comparison of six sigma and the lean organizations," The TQM Magazine, vol. 18, Issue 3, pp. 255-262, 2006.

[29]. Johannes, F., "State of the art concerning the integration of methods and techniques in quality management-literature review and an agenda for research," In ECIS, pp. 41, 2011.

[30]. Pfeifer, T., Reissiger, W., Canales, C., "Integrating six sigma with quality management systems," The TQM Magazine, vol. 16, Issue 4, pp. 241-249, 2004.

[31]. Marrs, F. O., Mundt, B. M., "Enterprise concept: business modeling analysis and design," Handbook of Industrial Engineering: Technology and Operations Management. New York: John Wiley & Sons, pp. 26-60, 2001.

[32]. Reid, R. A., Koljonen, E. L., Bruce Buell, J., "The Deming Cycle provides a framework for managing environmentally responsible process improvements," Quality Engineering, vol. 12, Issue 2, pp. 199-209, 1999.

[33]. Gidey, E., Jilcha, K., Beshah, B., Kitaw, D., "The plan-do-check-act cycle of value addition," Industrial Engineering and Management, vol. 3, Issue 124, pp. 2169-0316, 2014.

[34]. Henderson, K. M., Evans, J. R., "Successful implementation of Six Sigma: benchmarking general electric company," Benchmarking: An International Journal, vol. 7, Issue 4, pp. 260-282, 2000.

[35]. Kumar, S., Sosnoski, M., "Using DMAIC Six Sigma to systematically improve shop floor production quality and costs," International Journal of Productivity and Performance Management, vol. 58, Issue 3, pp. 254-273, 2009.

[36]. Kumar, N., Stern, L. W., Anderson, J. C., "Conducting inter organizational research using key Informants," The Academy of Management Journal, vol. 36, pp. 1633-1652, 1993.

[37]. Hailu, H., Kedir, A., Bassa, G., Jilcha, K., "Critical Success Factors Model Developing for Sustainable Kaizen Implementation in Manufacturing Industry in Ethiopia," Management Science Letters, vol. 7, pp. 585-600, 2017.

[38]. Hailu, H., Taddese, F., Tsegay, K., Jilcha, K., Hailu, T., "Relationship Between Kaizen Philosophy and Organizational Performance Empirical Investigation: a Case of Ethiopian Manufacturing Industries," European Online Journal of Natural and Social Sciences, vol. 9, Issue 4, pp. 735-751, 2020.

[39]. Field, A., "Discovering statistics using SPSS," Sage publications, 2009.

[40]. Valmohammadi, C., "Identification and prioritization of critical success factors of knowledge management in Iranian SMEs: An experts' view," African Journal of Business Management, vol. 4, Issue 6, pp. 915, 2010.

[41]. Bryman, A., Cramer, D., "Quantitative data analysis with SPSS 12 and 13: A guide for social Scientists," Psychology Press, 2005.

Appendix A

Tooling (the input box, toolbox and output box) the proposed integrated CI framework

	Input box	Toolbox	Output box
Monitoring and Controlling	**Phase V: Verifying, Sustaining and Complete CI Transformation**		
	CI experts, agents, juries, a set of rules, Change requirements, Progress sharing session	Judgment of experts, agents and juries, experience sharing session, Change control, Documentation of ICI practices	Appraisal report, Best performers, Review of organizational CI lessons learned documents (best practice document), Standardized CI practices, CI scope expanding decision, and Final project report
	Phase IV: Performance Evaluation or Stabilization		
	Questionnaire, CI experts, CI agents, manufacturing industries annual report, Progress sharing	CI Expert judgment. CI Agents judgment, Histogram, Scatter plot, Control charts, Overall Equipment Effectiveness	Preliminary evaluation report, Progress report, implementation decision, Identified operation, innovation & business results
	Phase III: Implementation and monitoring		
	CI implementation teams, CI implementation plan, Selected unit for implementation, CI experts, CI agents/Institutions, Employees engagement, Updating improvement actions, Progress sharing session	Expert judgment, CI implementation guidelines, Determined CI (basic and/or advanced) practices, Communication tools and skills, CI agents/Institutions judgment, Unique & common practices, 5S, Poka1-yoke, QC Story, Task Achievement, ISO 9000 Series, OEE, benchmarking, best practice)	CI pilot project, Standardized work activities, CI lessons learned of the pilot project, Progress report
	Phase II: Implementation Design including strategic planning		
	Organizational assets, Questionnaire requirements, Work activities details, Customer requirements, Link with suppliers, Organization structure, Effective communication, Managers involvement, Performance data, Process capability (CP), Customer management, Market analysis, Progress sharing session	Policy management, Set up performance measures, SWOT analysis, Value Stream Map, Gap analysis using sigma (δ) measure (DPMO), CP, Yield, Design of experiment , Cause and effect analysis, Pareto analysis, Voice of customer , SIPOC, Quality Function Deployment, Failure Mode Effect Analysis, Statistical process control, CI workshops and learning	Teams, Questionnaire of CI assessment, Documented current state gap, List of problems, Problems main root causes, Problems critical root causes, Problem elimination tools, Organizational CI expertise, Policy document with setup objectives, policies & strategies, Identified possible solutions, Identified customers & suppliers, Developed control plan, Progress report
	Phase I: Conceptualization		
	CI experts, CI agents, Previous CI index, documented practices, Problem profile, Customer requirements, Link with suppliers, Questionnaire requirements, Kick-off session, Progress sharing session	Expert judgment, CI knowledge, Training and education, Training games, Tuck man's model, CI preliminary analysis, Cause and effect analysis, Pareto analysis, Pre and post training assessment of trainees	Trained CI expert team, Trained & developed employees, Review of the lessons learned, Problem type list, Organizational CI practices, Questionnaire of CI assessment, Planning document, Trainees assessment result, Progress report

Role of Discrete Event Simulation in the Assessment and Selection of the Potential Reconfigurable Manufacturing Solutions

Mohsin Raza, Arne Bilberg, Thomas Ditlev Brunø, Ann-Louise Andersen, Filip Skärin

***Abstract*—** Shifting from a dedicated or flexible manufacturing system to a reconfigurable manufacturing system (RMS) requires a significant amount of time, money, and effort. Therefore, it is vital to verify beforehand that the potential reconfigurable solution will be able to achieve the organizational objectives. Discrete event simulation offers the opportunity of assessing several reconfigurable alternatives against the set objectives. This study signifies the importance of using discrete-event simulation as a tool to verify several reconfiguration options. Two different industrial cases have been presented in the study to elaborate on the role of discrete event simulation in the implementation methodology of RMSs. The study concluded that discrete event simulation is one of the important tools to consider in the RMS implementation methodology.

***Keywords*—** reconfigurable manufacturing system ; RMS; discrete event simulation; Tecnomatix Plant Simulation.

Mohsin Raza is with the University of Southern Denmark, Denmark (e-mail: raza@iti.sdu.dk).

Validating Condition-Based Maintenance Algorithms Through Simulation

Marcel Chevalier, Léo Dupont, Sylvain Marié, Frédérique Roffet, Elena Stolyarova, William Templier, Costin Vasile

Abstract**—** Industrial end users are currently facing an increasing need to reduce the risk of unexpected failures and optimize their maintenance. This calls for both short-term analysis and long-term ageing anticipation. At Schneider Electric, we tackle those two issues using both Machine Learning and First Principles models. Machine learning models are incrementally trained from normal data to predict expected values and detect statistically significant short-term deviations. Ageing models are constructed from breaking down physical systems into sub-assemblies, then determining relevant degradation modes and associating each one to the right kinetic law. Validating such anomaly detection and maintenance models is challenging, both because actual incident and ageing data is rare and distorted by human interventions, and incremental learning depends on human feedback. To overcome these difficulties, we propose to simulate physics, systems and humans – including asset maintenance operations – in order to validate the overall approaches in accelerated time and possibly choose between algorithmic alternatives.

Keywords**—** Degradation models, Ageing, Anomaly detection, Soft Sensor, Incremental learning.

I. Introduction

NOWADAYS, digitization and Industrial Internet of Things (IIoT) make it extremely easy to collect a vast amount of data concerning electrical assets during their operational life in real time conditions on customer plants. It allows users to get information on major environmental and usage conditions for products in real situations, together with data on observed failures.

Collected data can be used to learn normal behavior models of assets. Such models are relevant to detect anomalies, characterized by a statistically significant deviation from the so-called "current normal situation". While this kind of anomaly detection method is suited for short term asset monitoring, it is not appropriate for long term degradation trends. To capture such trends, first principles degradation models can be used (i.e. physical degradation models such as Arrhenius law); they complete the monitoring system, to get a global view on asset health. Both approaches are detailed in [1].

Validating such approaches represents a significant challenge, as only little real incident data is available from the field while degradation spans across several years of normal operation. Long-term degradation data collection is impacted by actual maintenance operations that may not have happened at optimal time. Furthermore, in the Machine Learning-based approach, models are created from collected data using incremental learning strategies. During this process, useful data may be accidentally dropped, while fault data may be mistakenly injected as "normal" in the update steps.

Simulation and validation of maintenance models is first presented in section II; incremental learning-based short-term virtual sensor models are addressed in section III. We finally conclude in section IV.

II. Long-Term

In this part, we refer to a long-term approach to compute the ageing of assets with respect to time, environmental conditions, usage conditions, and maintenance operations. These ageing models can be used to define appropriate time-based maintenance.

In the following sections we will present how these models can be used to simulate alternate system designs and maintenance scenarios, e.g. condition-based maintenance.

A. Proposed approach

Each asset to monitor is decomposed in sub-assemblies, all associated with one or several degradation modes [1]. These degradation modes act in competition for each concerned sub-assembly. At each point in time, the most impacting degradation mode is selected for each sub-assembly, and the most impacted sub-assembly is used to compute the consumed lifetime. Maintenance operations may replace the currently most impacted sub-assembly with another one, or change the current most impacting degradation mode with another one.

Fig. 1 illustrates a simulation run for a given asset, showing the Consumed lifetime (%) over time. Each spike is due to a maintenance operation of replacement of a sub-assembly,

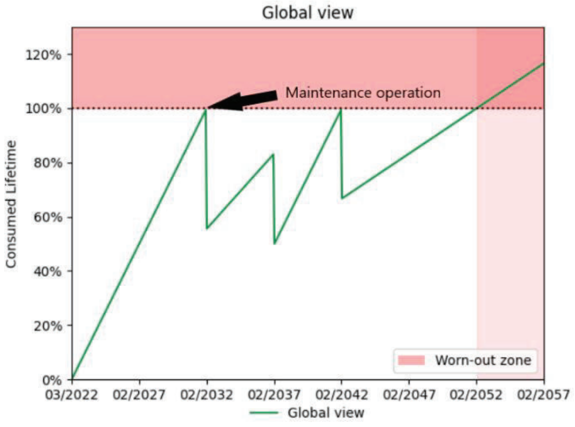


Fig. 1 Long term simulation

All authors are with Schneider Electric Industries, 160, avenue des Martyrs, 38000 Grenoble, FRANCE (e-mail: first.last@se.com).

which resets the ageing value. The first maintenance operation is done on the most impacted sub-assembly. After this event, the consumed lifetime is reset to ~60% and the degradation speed (slope) changes, as the most impacted sub-assembly is now a different one (the first one has been maintained).

What-if analysis is defined in the literature as a data-intensive simulation whose goals are to inspect the behavior of a complex system [11]. Initially focusing on scalar data, what-if scenarios nowadays embed more and more timeseries data to improve precision in the analysis [8]. Two kinds of what-if analysis can be considered:

- *Sensitivity analysis*: by generating a great number of scenarios, we can test the sensitivity of the ageing models to refine them. This analysis allows verifying and quickly modifying the models, so as to be as reliable as possible;
- *Scenario analysis*: it is also possible to generate realistic scenarios, to match reality as much as possible. Doing so, different choices can be simulated and confronted in order to find the best solution.

In this study, we focus on scenario analysis to explore two dimensions: alternate usage contexts (B) and alternate maintenance scenarios (C).

B. What-if analysis: alternate usage contexts

By modeling various scenarios, we can simulate the ageing process in a long-term view.

Our simulations of environmental conditions are based on weather simulation data [4] to provide the most realistic environmental scenarios. Naturally, these scenarios depend on the location of the customer site, together with the future usage of our products. We also have to link all the influencing factors and the customer events. For example, if a customer wants to add an air conditioner, it is possible to simulate the direct impact on environmental entries. In this case, the temperature remains constant; the humidity could be also static, but there are possible drawbacks like a raise of dust due to a new ventilation. So, with a realistic simulation, we can assess the impact of the air conditioner to advise the customer about this choice. An example of that case is provided in Fig. 2.

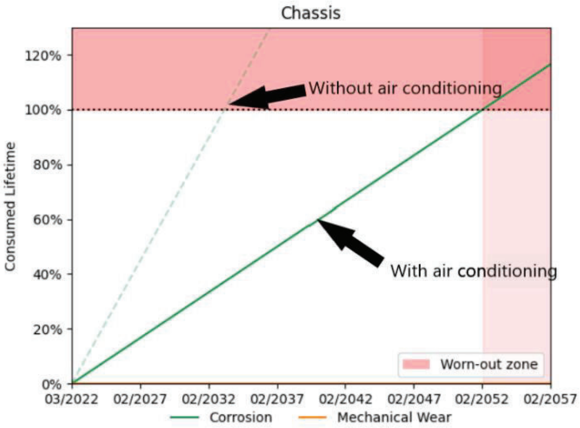
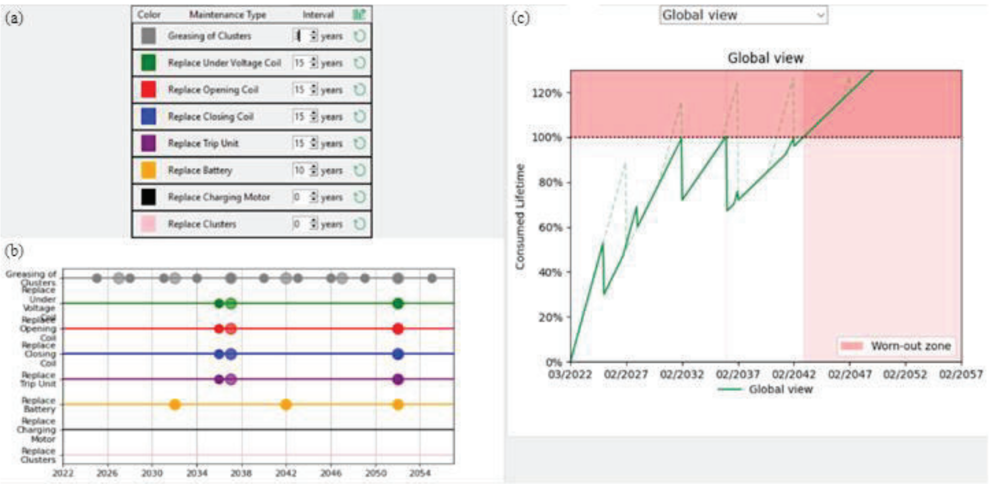


Fig. 2 Air conditioning impact

C. What-if analysis: alternate maintenance scenarios

Concerning maintenance, what-if scenarios allow to optimize the frequency of maintenance plans that can be adapted to each customer.

Most often, maintenance plans are extracted from maintenance guides, based on fixed periodicity recommendations [9]. The described approach enables the generation of adequate matches to plan maintenance operations, and advise our customers about the risk taken by withholding manufacturer maintenance. This is possible by simulating the life operations of the customer asset, and then performing maintenance on a subcomponent only when it reaches its end of life by comparing two scenarios : the standard maintenance plan and the custom one depending on ageing. We have compared these two scenarios in Fig. 3, where the effects of maintenance actions are reflected on the curve by sudden drops in lifetime consumption.



Color	Maintenance Type	Interval
■	Greasing of Clusters	1 years
🟩	Replace Under Voltage Coil	15 years
🟥	Replace Opening Coil	15 years
🟦	Replace Closing Coil	15 years
🟪	Replace Trip Unit	15 years
🟧	Replace Battery	10 years
⬛	Replace Charging Motor	0 years
■	Replace Clusters	0 years

Fig. 3 Maintenance impact

Fig. 3 (a) shows the maintenance periodicity selector used to configure the maintenance scenario. (b) illustrates how the selected scenario differs from a reference scenario and provides maintenance periodicity fine-tuning capabilities. Finally, (c) provides a comparison of the selected scenario with the reference one in terms of ageing.

D. Field deployment

The previous section on what-if scenarios described how simulation helps the validation step, in terms of both ageing simulation and maintenance scenario.

These simulation processes allow to transform traditional maintenance practices into a condition-based maintenance approach. Indeed, maintenance manufacturer periodicity recommendation described in maintenance guides are useful to provide an overview of the equipment lifetime. However, depending on type of segments (marine, mining-minerals-metals, oil and gas, healthcare, food & beverage), constraints are not the same. Two examples are described below:

- Marine customers have the specific constraint to make maintenance action when boats return to dock. Being able to estimate the ageing of the assets and consequence of maintenance actions facilitates the anticipation of required actions, ensuring confidence in the asset behavior waiting for the next maintenance while cruising.
- Regardless of the type of segments, the number of equipment is not the same from one customer to another. When the number of equipment is high, this kind of simulation helps to manage the fleet maintenance management by anticipating consequence of maintenance rescheduling for part of the equipment.

Therefore, combining manufacturer periodicity recommendation and simulation process allows to adapt the maintenance actions to each customer by adding more flexibility, more anticipation of end-of-life, and making timely maintenance. This condition-based maintenance approach makes operations more efficient while making business more resilient and sustainable [12].

With Schneider Electric, this approach is used by customers from different segments and for different kinds of assets, on low voltage or medium voltage domains. In the coming two years, we expect to grow by 200% the number of customers asking to transform traditional maintenance practices into condition-based maintenance ones. By increasing the number of field experiments, we will continuously challenge and improve the current models of ageing estimation, leveraging the benefit to link simulation and validation by closing the loop from customer, fostering its data to improve the models.

III. SHORT-TERM

Section II showed how what-if simulations can help decide and validate some maintenance actions to reduce long-term failures, or increase expected lifetime. This section focuses on short term analysis to detect unexpected failures. We first remind the proposed approach using machine learning model, then describe the incremental learning strategies, and finally illustrate with experimental results.

A. Proposed approach

Virtual sensors are Machine Learning-based techniques used for short-term anomaly detection [1]. Virtual sensors predict various industrial data based on usage and environmental conditions. Such idea of “virtual sensor”, also known as “soft sensor”, is not new and has been described in many ways in the literature and implemented in industrial products (see [2] for a review). A virtual sensor can predict a category (classification) or a quantity (regression). Besides their use for better process control, virtual sensors are used to tackle many other problems, such as back-up of a real sensor, what-if analysis, sensor validation, and fault detection and diagnosis.

The latter consists typically in monitoring a statistically significant deviation between the actual monitored data and the learnt reference provided by the virtual sensor. A usual way to perform this step is to model the prediction error (so-called residuals) and use this model to detect a statistical outlier [5]. Fig. 4 illustrates how such Virtual Sensor Fault Detection (VSFD) technologies can be assembled to create an adaptive temperature monitoring solution.

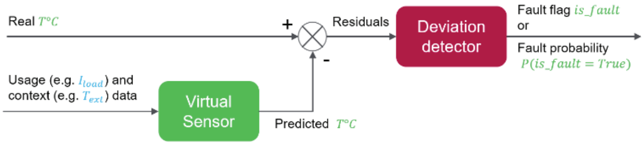


Fig. 4 Virtual Sensor Fault Detection Principle

B. Incremental learning framework

1) Context

Virtual sensor models are learnt on historical data capturing normal process operation. Such historical data spans over a limited period of time and therefore may describe a limited set of process operations. New data may contain unseen normal samples representing other or new aspects of the process. This notion known as *concept drift* is affecting the predictive accuracy of models over time. *Adaptive* Soft (Virtual) Sensors approaches [13] propose to update models using the new samples, or fully retrain models with the augmented dataset. Advanced sample selection and weighting techniques, as well as ensemble methods, are typically used to cope with challenging drifts [7]. Finally, recent uses of deep learning models to perform such tasks highlighted a specific issue known as the stability-plasticity dilemma, leading in extreme cases to *catastrophic forgetting* [10].

More general frameworks known as *Online*, *Sequential*, *Incremental*, *Lifelong*, or *Continual* learning [3] have recently emerged to describe applications where multiple aspects of the problem evolve with time. While the vast majority of work focuses on deep learning models for image classification, some papers mention regression tasks [6]. Adaptive virtual sensors can be seen as a particular case of incremental learning with a single task (regression), incremental on the data-domain.

Based on this review, we can delineate three general concerns in incremental learning regarding updates:

- *When* to update: model updates can be set to run periodically (e.g. weekly), or triggered whenever novelty is detected in the samples by a novelty detection model [14]. An alternate approach is to store novel samples in a buffer and update the model when maximum capacity is reached [6]
- *What* samples to integrate: too large new sample batches may be sub-sampled; under-represented class labels may be boosted with sample generation techniques; etc.
- *How* to inject the new samples: special update steps can include sample weighting for example

Most of the literature on incremental learning considers all samples to be safe for reinjection and is concerned with how to inject them. Some applications of adaptative virtual sensors are concerned with *what* data to reinject ; they tackle the issue using specific models and architectures [14]-[15]. These studies focus on classification models; besides they do not try to evaluate and compare different update strategies – but implement one and assess its performance.

There is no literature to date, to the best of our knowledge, comparing incremental learning strategies for regression virtual sensors for fault detection tasks (III.A). Process faults arising in operational data are an extreme case of outliers or concept drift: the statistical properties of the data during fault periods become different from the normal one used in the training set. However, as opposed to the usual concept drift handling approaches, the updated model should be protected from deviations so that it can still detect such faults in the future. In other words, new samples to integrate in the model should be carefully curated to eliminate abnormal samples and preserve its aim at modeling normal behaviors.

2) General Framework

An initial model M_1 is learnt on an initial dataset D_1. D_1 is considered safe: it only contains normal (non-fault) data. The following steps are then performed in sequence:

1) Use model M_1 to predict the target variable and detect faults during next period (dataset D_2)
2) Based on step 1 outcome, filter dataset D_2 to keep only "safe" samples $D_2^{'}$
3) Update model M_1 with $D_2^{'}$ to create model M_2
4) Iterate: repeat step 1-3 with new model M_2 and next period D_3

Fig. 5 below illustrates this procedure. The curation step (2) is indicated with a star (*).

Note that this procedure describes online learning-style update steps. It can be easily adapted to include full model retraining instead of model updates. Associated alternate Step 3 becomes:

3') Append $D_2^{'}$ to previous dataset D_1. Train new model M_2 using the merged dataset.

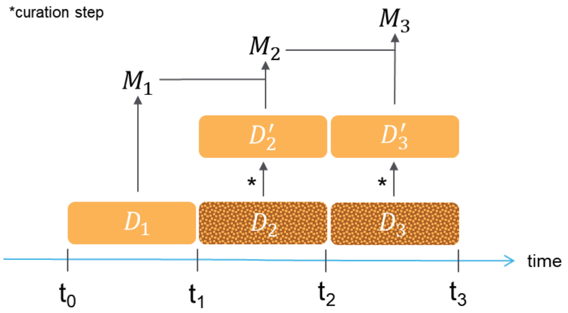


Fig. 5 Incremental learning framework

3) Considered implementation

Within the framework described in III.B.2), we consider the particular implementation below. An initial training period of 4 weeks is considered for D_1. Incremental learning is done on a weekly calendar basis. The curation step (2) is composed of two sub-steps:

- *Automated faults filtering*: even when the model predicts a fault, some samples are automatically re-tagged as normal.
  - *Level 1*: previously unseen (out-of-training range) data is re-tagged as normal,
  - *Level 2*: in addition to level 1, transient faults and faults tagged as unrealistic by an expert rule are re-tagged as normal. Such an expert rule includes physics-driven concerns, for example in some contexts an "under-heating" fault is not likely to be an actual fault but rather a model prediction error.

- *Faults validation*: faults are then presented to a human operator (expert). This operator is in charge of providing the final label: it can either "confirm the fault" or "discard the fault". Discarded faults are re-tagged as normal.

Finally, models are fully relearnt (3'.) instead of being updated (3.), so as to eliminate suboptimality issues related to partial updates. The size of the datasets is small and the considered model simple, thus full retraining is not prohibitive; and neither security nor storage are issues here. The overall process is described in Fig. 6.

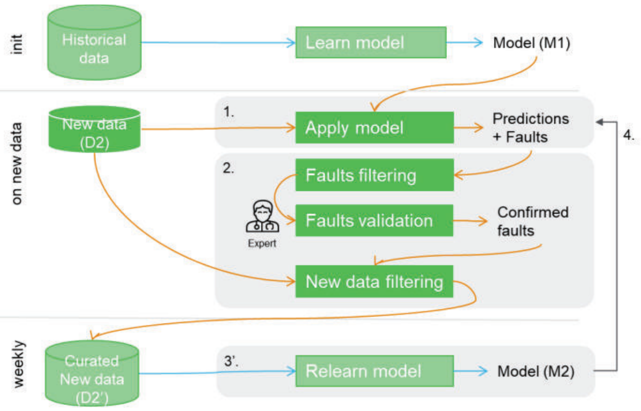


Fig. 6 Virtual Sensor Fault Detection orchestration

C. Validation using simulation

Validating the adaptive VSFD orchestration presented in III.B.3) is challenging, as only a little real incident data is available from the field. To overcome these difficulties, we propose to simulate the entire procedure on realistic datasets. This allows us both to validate correct behavior (overall feasibility) and to choose between algorithmic configurations.

1) Annotated datasets

Datasets injected in the simulator contain actual measurements (a multivariate timeseries) recorded from the field during several months of process operation. Each dataset is annotated: it is associated with zero, one or several *fault periods*. A fault period consists of a start and end date, and corresponds to a real incident observed on the field. Fault periods are usually defined as broader than the actual fault, because it is assumed that unobserved fault samples may already be present before the fault has been detected by a human operator. Also, logbooks from the field may be imprecise and contain vague fault period definitions.

2) Algorithm configurations

We consider eight alternate virtual sensor configurations related to the size of sliding window used: 1h, 2h, 3h, 6h, 9h, 10h, 12h, and 24h. For a given virtual sensor configuration, we consider 4 alternate fault detection threshold multiplicator (see [5]): 2, 2.5, 3, and 3.5.

3) Update strategies

Although most steps described in previous section III.B.3) present no particular technical difficulties, a challenge comes with simulating the human expert contribution happening in step 2 (fault validation). Besides, the *Level 2* automated faults filtering step may mistakenly re-tag actual fault samples as normal ones. We define below three strategies to simulate the fault filtering and validation process: *conservative*, *realistic*, and *oracle*.

Conservative strategy: the conservative strategy represents a low-risk scenario. Only *Level 1* fault filtering is active, and all other faults keep their label. *Level 2* is inactive and the expert always confirms faults. In other words, only samples not tagged as fault by the VSFD and samples out-of-range of the training space are considered normal and used to relearn the model. In this scenario, a minimal volume of new data is reinjected for learning every week.

Realistic strategy: this strategy is quite similar to the conservative one. This time, both *Level 1* and *Level 2* fault filtering are active, but the expert cannot discard faults. In other words, in addition to the conservative strategy, samples representing non-physically valid or transient faults are reinjected too. In this scenario, a medium volume of data is reinjected for learning every week.

Oracle strategy: in the oracle strategy, both *Level 1* and *Level 2* fault filtering are active, and the expert is omniscient and knows exactly when actual faults occur. In other words, in addition to the realistic strategy, samples detected as fault and not re-tagged by the filtering are automatically re-tagged as normal by the expert if they fall outside the actual fault period. In this scenario, a maximal volume of new data is reinjected for learning every week.

The conservative and oracle strategies have opposite behaviors and serve as representative bounds of the reality. In the conservative and realistic strategies, the expert is not confident on its abilities to recognize faults, while in the oracle scenario, it never makes mistakes. In the real world, experts are somewhere between realistic and oracle. They can adapt their behavior to particular customers and assets, as well as use their expertise and analysis skills on the measurement timeseries to discard false alarms. Simulating such an adaptative strategy for experts is out of scope of this study.

4) Assessment metrics

We introduce the following performance metrics to evaluate how well a given simulation run has succeeded in terms of fault detection.

Each sample is tagged as *False Positive* (FP), *True Positive* (TP), *False Negative* (FN) or *True Negative* (TN) depending on whether its label after the *Automated faults filtering* step matches the ground truth label (True/False) and if it is fault (Positive) or normal (Negative).

The *fault detection indicator* FD is defined as a boolean/dummy variable, equal to 1 when at least one fault was detected within the fault period, and to 0 otherwise.

The *fault periods coverage* FPC (a.k.a. *sensitivity* or *recall*) is defined as the ratio between the number of samples tagged as fault in the fault periods and the total number of samples in the fault periods.

The *false alarm ratio* FAR (a.k.a. *false positive rate*) is defined as the ratio between the number of samples tagged as fault outside of the fault periods and the total number of samples outside the fault period.

$$FPC = \frac{TP}{TP + FN} \qquad FAR = \frac{FP}{FP + TN}$$

5) Experimental results

We use data from three customer assets: two low-voltage panels and one medium-voltage panel. Inside a panel, a data set may be representative of different sub-assemblies: cable, busbar or withdrawable circuit breaker connections. Datasets span several months and contain either zero or one annotated fault period, that may be approximate (see III.C.1). Table I describes the five datasets used in this study.

We run the procedure described in (III.B.3): for each of the five cases, eight algorithm configurations, four fault detection threshold multiplicators, and three update strategies; giving in total 480 alternatives to evaluate. The *FPC* and *FAR* metrics are computed on each alternative. They are then averaged across all datasets, so that each model configuration is associated with a single pair of metrics.

Fig. 7 presents the results obtained in this experiment. For each of the considered update strategy (subplots (a), (b), (c)), the *FPC* (x-axis) and *FAR* (y-axis) are displayed for all alternatives algorithm configurations. Alternatives are grouped by sliding window size (colored line), where each group

TABLE I
DESCRIPTION OF USE CASES (15 MINUTES DATA SAMPLING RATE)

Case	Customer	Asset	Sub-assembly	Fault period	Dataset duration	Observations
1.1	1	LV Panel	N° 1	01/03/2019 – 15/04/2019	04/06/2018 – 08/03/2021	96,799
1.2	1	LV Panel	N° 2	No fault	04/06/2018 – 08/03/2021	96,799
2.1	2	LV Panel	N° 1	No fault	10/07/2019 – 11/03/2020	23,497
2.2	2	LV Panel	N° 2	01/10/2019 – 31/10/2019	10/07/2019 – 11/03/2020	23,497
3.1	3	MV Panel	N° 1	01/09/2020 – 30/09/2020	14/11/2019 – 24/03/2021	47,620

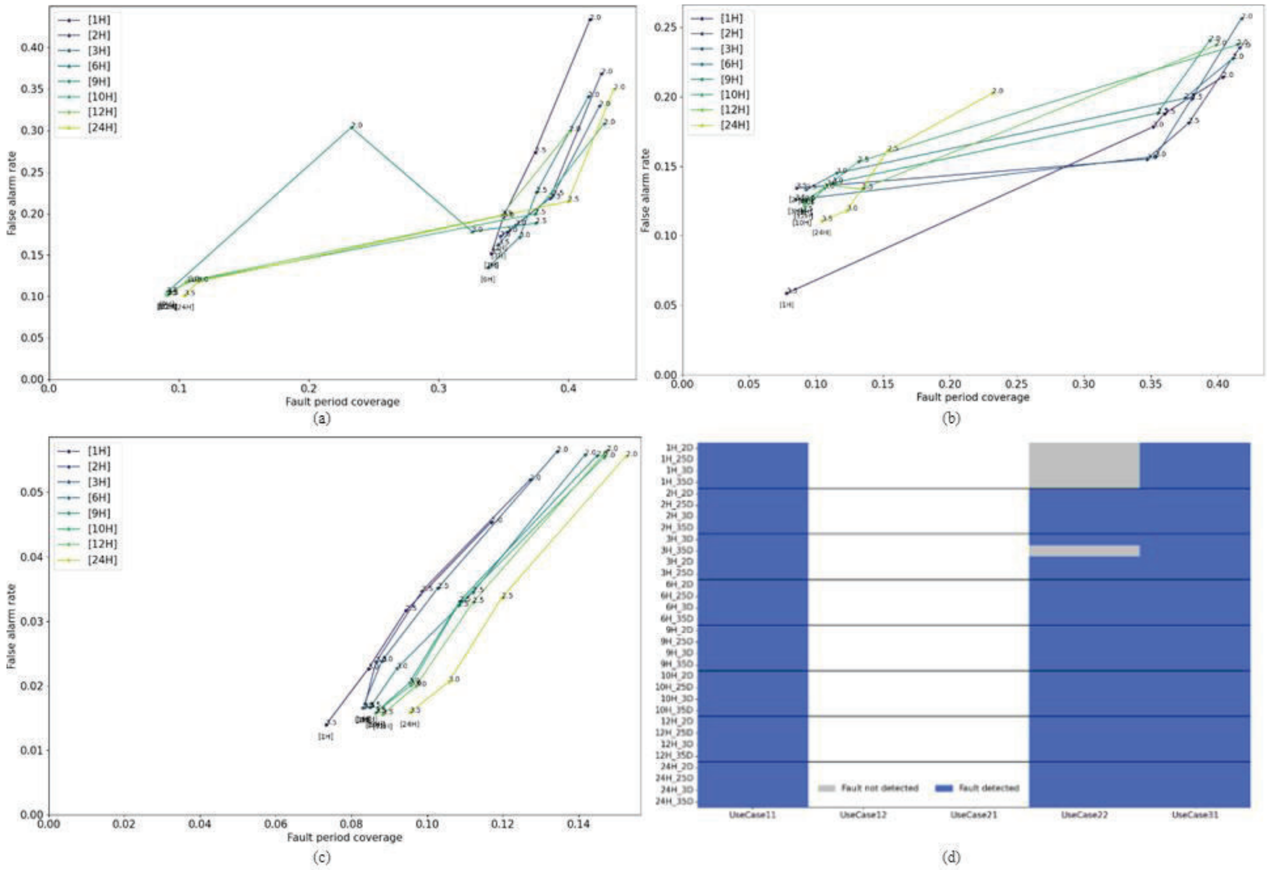


Fig. 7 Results for the Conservative (a), Realistic (b) and Oracle (c) strategies.
Fault detection heatmap (d)

contains results for the 4 fault detection multiplicator values. In addition, a fault detection heatmap for FD is computed for each strategy; however since all lead to identical results a single one is presented (d).

From the fault detection (*FD*) heatmap, we can first discard configuration [1H] (models using a sliding window size of 1h): whatever the fault detection multiplicator, none was able to detect the fault in use-case 2.2. This is also the case for configuration [3H] with multiplicator 3.5.

From the scatter plots, we see several trends. First, the amount of false alarms (*FAR*, y-axis) decreases as the strategy moves from Conservative (10-44%) to Realistic (5-25%) and finally Oracle (1-6%). This tends to confirm the role of a good expert feedback loop in the quality of incremental learning.

Concerning the fault prediction coverage (FPC, x-axis),

- In the *Oracle* scenario (c), all models have similar values (~10%). This consensus seems to indicate that the *actual* fault period is much smaller than the one declared in the datasets. In addition, we see that large sliding windows have better coverage: they are closer to 10%, while small ones are close to 5%. For the rest of analysis below, we use 10% as the optimal coverage to reach and consider higher values as "erroneously better".
- With the *Realistic* strategy (b), we see two groups of models with similar values (left: ~10%, same as in Oracle, or right: ~40%). The left group ("correct") corresponds to high fault detection multiplicators values, while the right group corresponds to low ones. In this scenario, the results

therefore tend to indicate that all models can be moved to the left group simply by increasing the fault detection multiplicator.

- With the *Conservative* strategy (a), the same two groups are visible, with the same values. However this time, models with small sliding window sizes (6h or less) are not able to move to the left group, even when the fault detection multiplicator is increased. It seems to indicate that to tackle this worst-case scenario, large sliding windows are preferable. We also note on this chart the presence of an outlier point.

Overall these results highlight a better stability of models using large sliding windows (>6h) and a large fault detection threshold multiplicator (3.5), even when the worst-case Conservative strategy is used. Adding *Level 2* filtering (III.B.3) brings a lot of value in terms of false alarms rate reduction and models stability. Finally, since these results were obtained on only five datasets, two of which without fault periods, it might be relevant to pursue larger-scale analysis to see if these trends are confirmed.

IV. Conclusion

Condition monitoring of critical assets requires both a short-term and a long-term perspective in order to both detect sudden anomalies and slow degradation. At Schneider Electric we combine Machine Learning methods for the short-term, and First principle statistical methods for the long-term. Validating such approaches is challenging. Indeed most of the real world collected data captures normal behavior and does not span for a long enough period of time to capture the asset degradation.

In this paper we propose to use simulation to overcome these difficulties:

- Long-term degradation simulation requires to simulate the environment and usage over decades of product life. Simulating the environmental conditions (Temperature, Humidity…) is performed using both timeseries modeling techniques (e.g. ARIMA) and modifications of real timeseries captured from sensors. Simulating usage is performed using modifications of real load profiles, representative of the diversity of use cases. Finally, several maintenance scenarios can be simulated in order to benchmark their efficiency: current field practices, optimized maintenance, etc.
- For Short-Term Machine Learning-based approaches, we can use actual field data in the simulator as a few months/years are sufficient to see practical convergence and accurate fault detection. However the challenge is associated with simulating model incremental learning over time, especially when a human operator is supposed to confirm/discard faults. We propose to bound the space of exploration with a pessimistic and an optimistic human behavior simulator, as well as a “probable” median scenario.

References

[1] Chevalier, M., S. Marié, B. Boguslawski, M. Cercueil, F. Chupot, A. Vignon, and W. Youssef. 2021. “Combining First Principles and Machine Learning for optimal maintenance of electrical assets”. In *CIGI QUALITA 2021*. May 2021, Grenoble, France.

[2] Curreri F., G. Fiumara, and M. Xibilia. 2020. “Input Selection Methods for Soft Sensor Design: A Survey”. *Future Internet* vol. 12 , no. 6: 97. https://doi.org/10.3390/fi12060097.

[3] Delange, M., R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, ... and T. Tuytelaars. 2021. “A continual learning survey: Defying forgetting in classification tasks”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. doi: 10.1109/TPAMI.2021.3057446

[4] Crawley D. B. and K. L. Lawrie. 2019. “Should We Be Using Just ‘Typical’ Weather Data in Building Performance Simulation?”. In *16th IBPSA Conference*, 2019. https://climate.onebuilding.org

[5] Gao, T., B. Boguslawski, S. Marié, P. Béguery, S. Thebault, and S. Lecoeuche. 2019. “Data mining and data-driven modelling for Air Handling Unit fault detection”. In E3S Web Conf., Volume 111, *CLIMA 2019 Congress*. Bucharest, Romania.

[6] He, Y. and B. Sick. 2021. “CLeaR: An adaptive continual learning framework for regression tasks”. *AI Perspectives* vol. 3(1), pp. 1-16.

[7] Kadlec, P., R. Grbić and B. Gabrys. 2011. “Review of adaptation mechanisms for data-driven soft sensors”. *Computers & chemical engineering* vol.35(1), pp. 1-24.

[8] Kegel. L, M. Hahmann, and W. Lehner. 2017. “Generating What-If Scenarios for Time Series Data”. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management* (SSDBM '17). Association for Computing Machinery, New York, NY, USA, Article 3, pp. 1–12.

[9] Masterpact MTZ Maintenance guide, 2015 https://www.se.com/ww/en/download/document/0613IB1202/

[10] Parisi, G. I., R. Kemker, J. L. Part, C. Kanan and S. Wermter. 2019. “Continual lifelong learning with neural networks: A review”. *Neural Networks* vol. 113, pp. 54-71. ISSN 0893-6080

[11] Rizzi S. (2009) What-If Analysis. In: LIU L., ÖZSU M.T. (eds) *Encyclopedia of Database Systems*. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-39940-9_466

[12] Schneider Electric, 2021, https://www.se.com/ww/en/work/services/service-plan/ecostruxure-service-plan.jsp

[13] Souza, F. A., R. Araújo and J. Mendes. 2016. “Review of soft sensor methods for regression applications”. *Chemometrics and Intelligent Laboratory Systems* vol. 152, pp. 69-79.

[14] Carino, J. A., et al. 2018. "Fault Detection and Identification Methodology Under an Incremental Learning Framework Applied to Industrial Machinery," in *IEEE Access*, vol. 6, pp. 49755-49766.

[15] Yu, Y., Peng, M., Wang, H., Ma, Z., Cheng, S., & Renyi, X. 2021. “A continuous learning monitoring strategy for multi-condition of nuclear power plant”. *Annals of Nuclear Energy*, 164, 108544.

TiO_2/PDMS Coating with Minimum Solar Absorption Loss for Passive Daytime Radiative Cooling

Bhrigu Rishi Mishra, Sreerag Sundaram, Nithin J. Varghese, Karthik Sasihithlu

***Abstract*—** We have designed a TiO_2/PDMS coating with 94% solar reflectance, 96% IR emittance, and 81.8 W/m^2 cooling power for passive daytime radiative cooling using Kubelka Munk theory and CST microwave studio. To reduce solar absorption loss in 0.3-0.39 µm wavelength region, a TiO_2 thin film on top of the coating is used. Simulation using Ansys Lumerical shows that for a 20 µm thick TiO_2/PDMS coating, a TiO_2 thin film of 84 nm increases the coatings reflectivity by 11% in the solar region.

***Keywords*—**Disordered metamaterial, Kubelka Munk theory Passive daytime radiative cooling, Solar absorption losses.

B.R. Mishra, S. Sundaram, N.J. Varghese, and K. Sasihithlu are with the Department of Energy Science and Engineering, IIT Bombay, Mumbai, India. (email ids of authors in the mentioned order: bhrigu.rishi_mishra@iitb.ac.in, sreerag@iitb.ac.in, nithinjov@gmail.com, ksasihithlu@ese.iitb.ac.in)

Fired Wood Metal Casting

Abdelrahman Hagmusa Idriss

***Abstract*—** We were assigned in 2017 by the Sudanese Ministry of Industry to carry out maintenance and manufacture of large pumps belonging to the Rahad Agriculture Project. The work is divided into many sections according to the manufacturing processes. The pump has many parts, some can be made by fabrication, but the impeller belongs to the casting section. Since that time, I have been thinking of casting methods to give me top accuracy and high quality by using the facilities available in my country. I found that sand molding was so difficult, so I began to think more and more, and suddenly fired, wood casting had been running in my head. Fired wood casting is for wide and complicated pieces, like lost wax casting for small and complicated pieces. The idea is simple words is to make a wooden pattern same like the piece with all details and also make the gating system from wood and then cover the patterns with sand like sand casting and then fire the wood to give a cavity .and the other steps same like the steps in sand casting. The paper contains all details for explaining the idea, like making molding, firing the wooden patterns, cleaning the arches, and pouring the molten metal.

***Keywords*—** new metal casting, new method of casting by firing the wooden patterns, fired wood casting for wide and complicated pieces, new method for ciplicated, large pieces by removing the wooden pattern by firing.

Abdelrahman Hagmusa Idriss is with the Sudan (e-mail: abdohagmusa@gmail.com).

Gradient Overdrive: Avoiding Negative Randomness Effects in Stochastic Gradient Descent (SGD)

Filip Strzałka Urszula Markowska-Kaczmar

Abstract—This work aims to develop a new method that maximally reduces the phenomenon of scrabbling weights in modern Deep Neural Network architectures without losing positive generalization characteristics of SGD. The goal of the conducted experiments is to tune the proposed method called Gradient Overdrive (GO) and try to prove its effectiveness by comparison to similar state-of-the-art methods. The method aims at achieving steeper learning curves in the same training regimes. Though the method should mark by being computationally efficient, neither the experimental implementation ensures to be optimal nor is it in the scope of this work to optimize the technique in the domain of computation time.

Keywords—neural network training, SGD, MLP, convolutional network.

I. Introduction

Deep Neural Networks (DNN) are entering their second decade of rampant development with numerous use cases and wide-scale research interest. The most common way of their training is the Stochastic Gradient Descent (SGD) algorithm. Despite broad research interest, there has long been no strict and formal understanding of the reasons why SGD produces a beneficial result. Particularly in the matter of its intrinsic randomness and its influence on the quality of training in different use cases, especially nonconvex problems [4].

A thorough mathematical analysis of DNN training algorithms has been provided only recently, e.g., overall statistical error analysis [5], and formal generalization bounds of SGD for DNNs in the over parametrized regime [6]. Some novel approaches to accelerating gradient in SGD have also been proposed and shown to be robust to statistical error accumulation and inherent instability in applications with Least Squares Regression (LSR) [9]. However, the trade-off between generalization effects and slow convergence still remains an open issue, especially in times of rapidly changing state-of-the-art DNN architectures.

New benchmark records achieved by novel yet intuitive techniques show us that an intuition-oriented fresh look or even rethinking a previously known method is key to achieving better results.

A novel approach to methods focused around optimizing SGD can be observed in recent works, including rethinking of the alignment between those methods and hardware acceleration [11]. There is increasing interest in the study of SGDs randomness. Several techniques aiming to exploit its error characteristics were either newly presented or rethought.

A common idea in recent work in this field has focused on the random noise introduced intrinsically by SGD and its positive and negative results (namely, good generalization capabilities versus slow convergence). One of the approaches is to introduce anisotropic noise in the SGDs optimization formula intentionally. It is a way of improving the algorithm's regularisation effects and its ability to escape from sharp minima [12]. The authors show that trying to domesticate the randomness by introducing extrinsic random distributions into the training regime yields a faster convergence rate. Parameter variance reduction is another important idea present in many recent works. It has been applied to improving the efficiency of minibatch training by using stratified sampling [13] and shown to be able to improve the convergence rate significantly in a popular setup where minibatch training is employed to parallelize SGD.

The same inherent variance has also been blamed as a cause of slow asymptotical convergence in [14] where an unbiased method called Stochastic Variance Reduced Gradient (SVRG) was introduced to remedy this problem with satisfying results for smooth and strongly convex functions. The approach has also been proposed in biased alternatives, namely Stochastic Average Gradient (SAG) [15] (SAG) and Point-SAGA [16].

On the contrary, more recent work showed that the above variance reduction methods are hardly applicable to complex non-convex optimization problems encountered during training of modern DNNs [18]. The authors aim at exploring the causes of the failure of SVRG and related approaches in such cases and point at complications related to the concurrent use of data augmentation, batch normalization, and dropout. The authors conclude the need to find a different approach to reducing inherent variance and suggest employing techniques used in learning rate optimization algorithms. This approach would allow a hybrid method to cope with non-convex problems and interoperate with modern convolutional, recurrent and attention-based neural network architectures.

Learning rate optimizers themselves have also been proven effective in many general and specific tasks of training DNNs. A front-line of ADAM [19] and its variations [20] has recently flooded practical DNN applications as a most widely used first choice optimizer but also challenged researchers to propose new approaches. A recent work by Schmidt and al. [21] has benchmarked the most popular optimizers. It also showed that as so many variants have arisen in recent years, trying different optimizers is often the most effective way of searching for optimal hyperparameters, especially in resource-limited training scenarios.

A higher-level technique of a generic Random Learning Rate has also been employed and used on top of optimizers to improve classification tasks results. It has been proven to be the best strategy in small learning rate regime applications, yielding better regularization without extra computational cost [22]. This technique can also be viewed as another approach to exploiting SGD's randomness to achieve the highest quality and effectiveness results.

In this work, we present a novel approach, which seems

Urszula Markowska-Kaczmar is with the Wroclaw University of Science and Technology, Poland (e-mail: urszula.markowska-kaczmar@pwr.edu.pl).

to cull from the above ideas in the intuition of improving SGDs efficiency by avoiding the negative effects of its randomness while preserving its positive impact on generalization. Precisely, we focus on a phenomenon of scrabbling weights and propose a technique to avoid it by intentionally trimming gradient on a basis sampled from a random distribution. We see it as a way of reducing parameter variance during training and prove applicable alongside modern DNN improvement techniques, including dropout, batch normalization, and different kinds of data augmentation.

The paper is composed of four sections. The next one presented some basic concepts that we use in our method. Description of the method is described in section III. Section IV presents new activation functions that we used in the experimental part. In section V we show details referring to the experimental procedure. Performed experiments are described in section VI. Conclusions end the paper.

II. Fundamentals

In this paper, we consider layered, feedforward network architectures. The first one, where neurons are fully connected in neighbouring layers, is called the *Multilayer Perceptron* (MLP). The simplest one contains one hidden layer. It was used in the initial experiments during development of our method. Further, we will also describe the *Convolutional Neural Network* (CNN) that was applied to see whether the method scales. Generally, the MLP is composed of an input layer, one or many hidden layers and an output layer. The number of hidden layers and the number of neurons in each layer are hyperparameters of a network. The number of outputs depends on the problem solved by the network (classification or regression). For simplicity, in the description below we assume the simplest MLP network with one hidden layer. The signal processing in the network can be described as a sequence of matrix operations for a given pattern $\mathbf{x}$ given in Eq. 1

$$\mathbf{h}(\mathbf{x}) = f(\mathbf{x}\mathbf{W}^{(h)} + \mathbf{b}^{(h)}) \tag{1}$$

where $\mathbf{W}^{(h)}$ is a matrix of hidden layer's parameters, The paper is composed of four sections. The next one presented some basic concepts that we use in our method. Description of the method is described in section III. Section IV presents new activation functions that we used in the experimental part. In section V we show details referring to the experimental procedure. Performed experiments are described in section VI. Conclusions end the paper. $b^{(h)}$ is a vector of hidden layer's bias values and f is an activation function. There are many different activation functions, commonly used are: sigmoid, hyperbolic tangent, step function, ReLU, linear function, and softmax in the output layer for classification problems.

The network output is defined by Eq. 2.

$$\hat{\mathbf{y}}(\mathbf{x}) = f(\mathbf{h}(\mathbf{x})\mathbf{W}^{(o)} + \mathbf{b}^{(o)}) \tag{2}$$

where $\mathbf{W}^{(o)}$ and $\mathbf{b}^{(o)}$ are the matrix of weights and the vector of biases in the output layer, respectively.

Another type of neural network considered in this paper is the *Convolutional Neural Network* (CNN) [17]. In the convolutional layers, neurons are not fully connected. A given neuron is only connected to a defined subset of neurons in the subsequent layer. Moreover, weights assigned to these connections are shared between neurons of a single feature map of the convolutional layer. They are widely used in image processing because they learn to extract complex image features that serve the performed task the most. In each convolutional layer, one defines a *kernel* (or a *filter*) - matrix with assumed sizes, significantly lower than the image resolution. Filter values are tuned during training the network, they correspond to neurons' connections weights. A single convolutional layer may consist of multiple filters, each producing a separate feature map. While processing an image, convolutional filters are moved across the image stepwise by a constant number of pixels, and convolution operation is calculated. It defines the total activation for one neuron in a convolutional layer, Eq. 3.

$$y_{kij}(\mathbf{x}, \mathbf{W}) = \sum_{p=1}^{P}\sum_{q=1}^{Q} w_{kpq}\ x_{i+p,j+q} \tag{3}$$

where $\mathbf{x}$ is the input image, $\mathbf{W}$ denotes a matrix of convolutional filters (size: $K \times P \times Q$), K is the number of filters (feature maps), $P \times Q$ is the size of a single convolutional filter, and i, j are a single neuron's coordinates.

Similarly to MLP, convolution output values are processed by an activation function, and then pooling is applied. Its role is to decrease the size of a feature map. Usually it is implemented as max operation (*max-pooling*) or average (*average-pooling*) from the sliding window. Pooling allows a convolutional network to be more robust to small image rotations and translations.

A. Neural Network Training

Neural network training is performed as an optimisation task. We search for the cost (loss) function minimum. It defines the error the network makes approximating the target function $\mathcal{F}$. Most commonly used loss functions are: mean square error, binary cross-entropy, and categorical cross-entropy.

For complex, multilayered network architectures *Backpropagation* algorithm is in everyday use to train the network. It assigns the loss function gradient for the last network layer and then using the chain rule, computes the gradient value for the weights in the immediately preceding hidden layer. For a network with more layers, analogously, the gradient in the n-th layer is calculated by propagating the loss function gradient from $(n + 1)$-th layer. Weights are updated according to the following equation:

$$\mathbf{W}^n = \mathbf{W}^n - \mu\frac{\partial\mathcal{L}}{\partial\mathbf{W}^n} \tag{4}$$

where μ is the learning rate coefficient.

We can calculate the loss function gradient $\mathbf{g}^n$ for the weights (and biases) change in the n-th network layer using the following recursive definition.

$$\mathbf{g}^n = \frac{\partial\mathcal{L}}{\partial\mathbf{W}^n} = f^{n-1}\ \delta_n \tag{5}$$

where:

$$\delta_i = \begin{cases} \left(\mathbf{W}_{i+1}\delta_{i+1} \odot f_i'\right)^\top & i < L \\ \left(\mathcal{L}' \odot f_L'\right)^\top & i = L \end{cases} \tag{6}$$

where $\mathcal{L}$ is the loss function, L denotes the total number of layers (and the index of the output layer), $\mathbf{W}_i$ is the weight matrix between $(i-1)$-th layer and i-th layer, f_i is the activation vector in the i-th layer, $'$ denotes a value of the derivative (or gradient), and the operator $\odot$ is the Hadamard product, L is the assumed number of layers.

After one input pattern presentation we update the model weights and repeat the estimation for the next patterns in the hope that such approach will lead to the model convergence. This process, called Stochastic Gradient Descent (SGD) [1], makes it possible to quickly achieve high accuracy even in nonconvex highly-dimensional problems, in which the complexity of training examples and the number of model parameters is high, while the total number of examples is relatively low (e.g. In common use is mini-batch stochastic gradient , where in a single iteration, the network parameters are updated based on a small portion of the training dataset. Another improvement is using momentum. This method almost always helps accelerate gradients vectors in the right directions, thus leading to faster converging. We define momentum as a moving average of our gradients used to update the weights of the network.

B. Additional DNN-training techniques

Several common techniques used for improving DNN results are commonly used in practice. Long training may cause network overfitting; therefore, in practice, some other techniques are applied. One of them is *early stopping*. It relies on setting aside an additional validation dataset. In every iteration, it is used to monitor the cost value. The training is stopped when the validation cost increases. Another technique is data augmentation – a technique used to increase the number of training examples in NN training by altering input data in a way that it could possibly be altered in real-life scenarios. The next technique is dropout [8]. It can be viewed as a type of regularisation that causes training different architectures of the network. Batch normalisation [7] is also worth mentioning. It targets reducing Internal Covariate Shift defined as the change in the distribution of layer outputs due to model training. By keeping the distribution of each layer inputs to have zero mean and unit variance, the authors succeed in improving training speed. More formally, the method transforms layer inputs $\boldsymbol{x}$ of the current batch step in the following way:

$$\boldsymbol{x} \leftarrow \frac{\boldsymbol{x} - \mathrm{E}[\boldsymbol{x}]}{\sqrt{\mathrm{Var}[\boldsymbol{x}] + \epsilon}} * \boldsymbol{\gamma} + \boldsymbol{\beta} \tag{7}$$

where $E[\boldsymbol{x}]$ is the expected value vector and $Var[\boldsymbol{x}]$ is the variance vector, both estimated by all samples within the current batch (hence *Batch Normalisation*). Factors $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are trainable parameters.

Gradient Clipping (GC) [3] is a method of avoiding exploding gradient values, prototypically performed by simply limiting the gradient's norm. The standard GC adjusts the gradient before updating model parameters θ (weights and biases) as follows:

$$g \leftarrow \begin{cases} \lambda \frac{g}{\|g\|} & \text{if } \|g\| > \lambda, \\ g & \text{otherwise.} \end{cases} \tag{8}$$

where λ is clipping threshold, a hyperparameter which usually has to be tuned to the specific training scenario.

C. Learning rate optimizers

The before mentioned hyperparameter μ depicting learning rate (*lr*) ratio can be viewed as the strength of parameter updates in each step. Higher learning rates reduce the total number of steps needed for convergence, but reduce the accuracy as the steps become too large. On the other hand, smaller learning rates help to find more accurate loss minima, but make the training process comparatively slow.

In fact, the opportunities sought by researchers in developing learning rate optimizers resulted in a massive number of recent publications presenting different approaches to adjusting *lr* [21]. Several of them however, became more popular and remain within a some type of canon for use in many different problems. An example of a well-known *lr* optimisation algorithm is ADAM optimizer [19] The last procedure updates the network parameters θ (weights and biases) at each training step as follows:

$$\theta_t = \theta_{t-1} - \nu_t \left(\frac{\alpha m_t}{\sqrt{v_t} + \epsilon} + \lambda \theta_{t-1} \right) \tag{9}$$

where v_t and m_t are first and second moment exponential running coefficients defined as follows:

$$\begin{aligned} m_0 &= 0 \\ m_t &= \frac{\beta_1 m_{t-1} + (1-\beta_1) g_t}{\left(1 - (\beta_1)^t\right)} \end{aligned} \tag{10}$$

and

$$\begin{aligned} v_0 &= 0 \\ v_t &= \frac{\beta_2 v_{t-1} + (1-\beta_2) g_t^2}{\left(1 - (\beta_2)^t\right)} \end{aligned} \tag{11}$$

where $\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_1 = 0.999$, $\epsilon = 10^{-8}$, $\lambda \in \mathbb{R}$ (all can be viewed as hyperparameters and tuned to a specific scenario). Common variant of this method employs decoupled weight decay [20].

III. Method

In the standard SGD or minibatch-SGD training methods, we attempt to train the model by assuming that a single training sample (or a single minibatch of samples) can be used to estimate the general direction of model parameters towards global loss minimum.

A. Problem formulation

By using SGD, we introduce a type of regularisation, as we never fit the model to all available examples at once. By employing the assumption that a single batch would be a good estimation of a whole dataset, we can force a neural network to generalise well even when the number of training examples is low and the number of trainable parameters is relatively high. This, however, comes at a cost of weight errance in their way to the global minimum, which in fact is not guaranteed to be found at all. This errance is a side effect of SGDs intrinsic randomness and results in potentially longer training. We thus aim at reducing negative randomness effects in SGD while preserving its generalisation effects.

More formally, the goal is to provide such a method that having a certain reference SGD training configuration with:

- a ResNet-like, or a modern state-of-the-art CNN architecture, such as NFNet
- an image classification problem with a well-known benchmark dataset, such as CIFAR10 or ImageNet ISLRVC 2017
- any subset of modern techniques: minibatch training, learning rate optimisation, data augmentation, dropout, batch normalisation, gradient clipping

will result in an improvement of training by any of these means

- a significant reduction of the training length (i.e. making the network achieve the same loss results in fewer training steps)
- a significant increase in the classification TOP-1 or TOP-5 accuracy scores measured on training sets, i.e. achieve state-of-the-art results

by reducing a phenomenon of opposing gradient values that appear in subsequent training steps and lead to a potentially suboptimal training process.

B. Intuition behind avoiding opposing gradient values

The method presented in this work emerged from our initial experience with MLP training experiments, in which we observed a phenomenon we called scrabbling weights. We find that quite often, in the training of neural networks, the weights seem to ”scrabble” around some time-local mean values before they eventually turn to the right direction. From the point of view of several subsequent training steps, the model weights sometimes seem to respond too quickly to opposing gradient values and, in turn, increase their variance while not changing the mean value much. The phenomenon can easily be observed in trivial cases of neural networks composed of only several pairs of neurons. An example is presented in Fig. 1. It shows the case of a neural network with one hidden layer composed of 2 neurons and one output layer with two neurons. Here, we can observe scrabbling weights and biases in the hidden layer that lasts more than 12 000 iterations. According to our intuition, such opposing gradient values can cause SGD convergence to slow down while not necessarily being the sole mechanism that enables generalization and the convergence itself.

Scrabbling weights are not easily observable in larger networks, as obviously, the parameters cannot be so well visualized. However, we expect it to occur also in much larger networks and even in sophisticated training regimes that use data augmentation, dropout, batch normalization, semi-supervised learning, or gradient clipping. We find that these techniques can partially solve this problem. The same applies to using learning rate optimizers, as they do not aim precisely at preventing the weights to scrabble but to locally optimize the learning pace with the goal of better balancing between fast and accurate convergence.

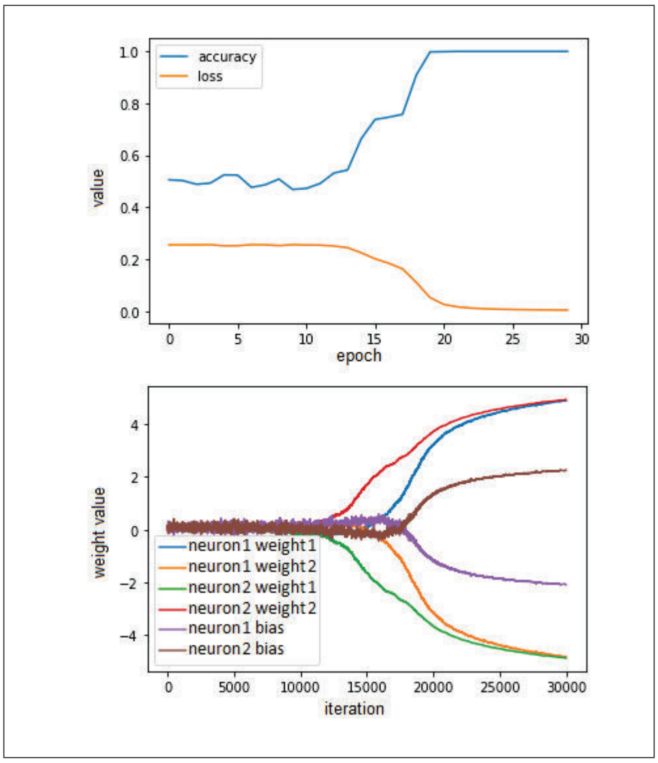


Fig. 1: On the top - plot of the training accuracy and loss function; bottom - hidden layer weight and bias values; the network trained on XOR problem examples. The weights scrabble in the first half of training time without taking up any particular direction. This time can be saved by preventing the phenomenon.

We find by multiple experiments that trying to reduce the phenomenon by avoiding opposing gradient values helps achieve higher training efficiency.

We name the proposed technique ”Gradient Overdrive” as an analogy to signal analysis theory, where overdrive is a result of trimming the amplitude of a signal to some constant maximal value.

This method can also be seen from the point of view of reducing gradient variance. In fact, reducing gradient values in some conditions is a common motif of many techniques to improve SGD’s efficiency. However, the key of our method is not to use continuous fractional changes to gradient values but to sample the binary decision of trimming it from the Bernoulli distribution with some adjusted probability.

C. Proposed solution: Gradient Overdrive algorithm

For a more straightforward explanation of the method and the intuition behind it, instead of matrix **W** or vector of biases **b** from now, we will refer to a single parameter (weight or bias value) k in the network.

We described it in the pseudocode within a standard mini-batch SGD training algorithm in Algorithm 1. We name Gradient Overdrive (GO), an operation performed in each training step, after computing the value of the loss function (line 2) and backpropagating it through the network (line 3), but before the lr optimization step. The algorithm is performed for each batch i. At this moment, in the standard SGD method, each model parameter should have a value $\widehat{g}_{k,i}$ scaled by lr optimizer algorithm and added to the parameter itself.

Algorithm 1: Gradient Overdrive pseudocode within a single training step

```
Data: Training examples of a next training (batch)
      step $i$
Result: Updated all $k$-th parameters $w_{k,i}$ in step $i$-th
1  perform forward pass;
2  compute loss value;
3  backpropagate to compute the gradients $g_i$;
4  for each single scalar parameter of the model $k$ do
5      estimate $EMA_{k,i}$ and $EMVar_{k,i}$;
6      compute overdrive probability $p_{k,i}$;
7      sample gradient overdrive flag $X_{GO_{k,i}}$;
8      if $X_{GO_{k,i}}$ equals $1$ then
9          overdrive gradient value $\widehat{g}_{k,i} = GO(g_{k,i})$;
10     else
11         do not overdrive, simply set $\widehat{g}_{k,i} = g_{k,i}$
12     end
13 end
14 having $\widehat{g}_{k,i}$ find $\Delta w_{k,i}$ by appropriate learning rate
     optimizer algorithm;
15 perform weight update: $w_k = w_k + \Delta w_{k,i}$;
```

The GO method is included in lines 4 to 13 (Algorithm 1). For all of the model's weight or bias (matrix or vector), for which we compute gradient $\mathbf{g}_i$ in step i, we have K single scalar parameters w_k (i.e., K is the number of all parameters in the network). We determine $\widehat{g}_{k,i}$ modifier for each k-th scalar parameter as follows (lines 8 to 12):

$$\widehat{g}_{k,i} = \begin{cases} \max[\mathrm{EMA}_{k,i}\ ,\ \ g_{k,i}], & \text{if } X_{GO_{k,i}} = 1, s_k = 1 \\ -\max[\mathrm{EMA}_{k,i}\ ,\ \ -g_{k,i}], & \text{if } X_{GO_{k,i}} = 1, s_k = -1 \\ g_k, & \text{if} X_{GO_{k,i}} = 0 \end{cases} \tag{12}$$

where:

$\mathrm{EMA}_{k,i}$ - exponential moving average of the scalar gradient value at step i
s_k - overdrive direction flag, for each k-th scalar parameter sampled once at the beginning of training from $\{-1, 1\}$ with equal probability
$X_{GO_{k,i}}$ - gradient overdrive flag - a random variable sampled from Bernoulli distribution with success probability p_i (line 6)

The intuition behind this equation is to effectively trim the gradient value above or below the estimated average ($\mathrm{EMA}_{k,i}$), direction depending on the random (but constant through training) flag value. The probability p_i is given by:

$$p_i = \max\left[0\ ,\ \ p_{max} \cdot \left(\frac{1}{K}\sum_{k=0}^{K} \frac{\mathrm{EMVar}_{k,i}}{\mathrm{EMVar}_{k,i}^{\max}}\right)^e - p_{cut}\right] \tag{13}$$

where:

$\mathrm{EMVar}_{k,i}$ - exponentially weighted moving variance of the scalar gradient value at step i
$\mathrm{EMVar}_{k,i}^{\max}$ - maximal value of EMVar_k from the beginning of training until step i
p_{max} - upper probability bound, GO hyperparameter
p_{cut} - probability cutout, a constant value to substract, GO hyperparameter
e - variance ratio influence index, GO hyperparameter

The intuition behind this equation is to gradually decrease the GO probability p_i through training exponentially to reduce the relative gradient variance of the whole model. The exponential index e, the maximal value coefficient p_{max} and the constantly subtracted value p_cut are hyperparameters that allow controlling the shape of the exponential-like curve, i.e., changing its steepness, maximal, and cutout value.

$$p_i = \max\left[0\ ,\ \ p_{max} \cdot \left(\frac{1}{K}\sum_{k=0}^{K} \frac{\mathrm{EMVar}_{k,i}}{\mathrm{EMVar}_{k,i}^{\max}}\right)^e - p_{cut}\right] \tag{14}$$

where:

Moments are estimated by exponential running average, according to Welford's online algorithm for computing mean and variance:

$$\mathrm{EMA}_{k,i} = \beta \cdot \mathrm{EMA}_{k,i-1} + (1-\beta) \cdot g_{k,i} \tag{15}$$

$$\mathrm{EMVar}_{k,i} = \beta \cdot \mathrm{EMVar}_{k,i-1} + (1-\beta) \cdot (g_{k,i} - EMA_{k,i-1})^2 \tag{16}$$

where parameter β is fixed at 0.8

Part of this technique (algorithm 1, line 5) is also used in popular Adam and AdamW learning rate optimization algorithms, which makes it favorable for using them cooperatively to save computation time (in line 14).

An essential issue towards applying these equations is to sample the directions of overdrive once before training randomly. Initial research showed that it is not favorable to change the overdrive directions later in training. This approach does not introduce direction bias for the whole network and ensures GO always ”forces” changes in the same way. As a result, we expect our method to be successfully combined with many other techniques used in modern neural network optimization.

D. Limitations of the presented method

One should notice that the expected value of a weight change differs from the classic SGD case as we don't compensate gradient trimming, which decreases the effective learning rate lr_{eff} from the point of view of its sole optimization:

$$\begin{aligned} lr_{\text{eff}} &= (1 - \mathrm{E}(p_i)) \cdot lr \\ &\leq lr \end{aligned} \tag{17}$$

where $\mathrm{E}(p_i)$ is expected value of gradient overdrive probability p_i among all model parameters during the whole training.

Possible compensation for that decrease was demitted in the initial stages of research, as it did not turn up to have positive effects on the convergence process. However, as our research only included experiments with Adam and AdamW optimizers (and initially simple mini-batch SGD with momentum), it can be considered a further study basis.

Another limitation of GO is its sensitivity to activation function characteristics. In particular, we find that not all of the traditional activations work well with the method. In research, we first test the most popular ones, including Sigmoid, Tanh, Hardswish, and popular variations of ReLU. We then provide several new activation functions based on intuition from initial experiments. The new activations are defined in the next section. However, it is noticeable that the search for optimal activation to work best with GO was rather rough than exhaustive.

The same applies to the method's hyperparameters, whose number (3) should profitably be reduced in further work. Heuristic approaches can be used in further study to solve these problems eventually.

IV. New Activation Functions Proposed

In our initial experiments, we found that constant slope (first derivative) values of the activation functions work worse with GO than those featuring curves, i.e., *logistic sigmoid* or *hyperbolic tangent*. Unfortunately, the two suffer from the vanishing gradient phenomenon and are more and more rarely used in deep neural networks. For this reason, we propose several new activation functions that combine ReLU with sine trigonometric function to find an activation that addresses both problems.

The intuition for creating new activations was based on the fact that the function that experienced the highest increase rates with GO was Sigmoid, which is hardly applicable to deep convolutional networks due to the vanishing gradient problem. Below we present their definitions along with sample plots shown in figure 2.

a) Sintoix:

$$f(x) = \sin x + 2x \tag{18}$$

b) ReSintoix:

$$f(x) = \begin{cases} \sin x + 2x, & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{19}$$

c) ShiftHalfSintoix:

$$f(x) = \begin{cases} \sin(x + \pi) + 2x, & \text{if } x \geq -\pi \\ 0, & \text{otherwise} \end{cases} \tag{20}$$

d) LeakySintoix:

$$f(x) = \begin{cases} \sin(x + \pi) + 2x, & \text{if } x \geq -\pi \\ \frac{1}{100} \cdot \sin x, & \text{otherwise} \end{cases} \tag{21}$$

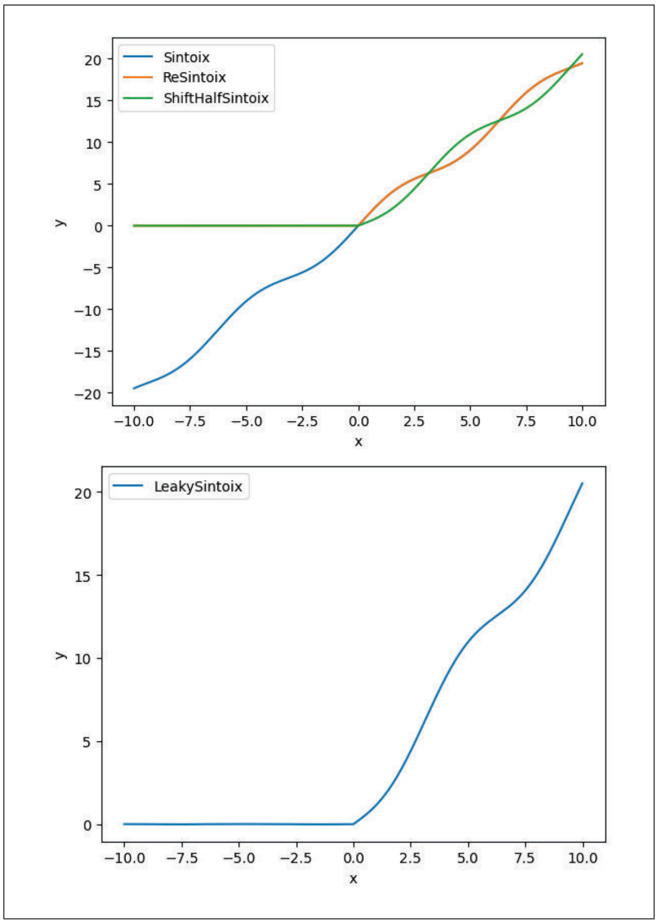


Fig. 2: Example plots of the newly proposed activation functions.

V. Experimental Setup

The general goal of experiments was to tune and evaluate the method's performance versus control cases by measuring the quality and efficiency of training CNNs of various architectures, sizes, and regularisation techniques, in common classification tasks. In the experimental research, we explore the influence of the network activation function and the method's hyperparameters on its properties. We consider a small (6-layered) convolutional network, 15 epochs of reduced CIFAR10 dataset, and a larger network (20-layer ResNet). The results are compared to the small network from the previous step with the same subset of hyperparameters found earlier. We also performed experiments to evaluate the method's qualitative and quantitative results on a larger model with modern regularisation techniques (EfficientNet) on the CIFAR100 dataset with data augmentation (random crop, random flip).

a) Experimental procedure:

For all of the conducted experiments, we took considerable caution in the matter of their reproducibility, i.e., the experimental procedure followed these assumptions:

- when run multiple times with the same parameters, random seed and on the same hardware setup, every single run of a training experiment should have the same result in terms of the model outputs on every step of training, and thus in the values of used loss metrics of any type,
- in the experiments comparing the results of different neural network architectures, the same initial weights and the same order of serving training cases should be retained..

To improve the statistical reliability of the results and ensure all experiments were treated equally, each particular experimental scenario (i.e., with given values of the variable parameters) was repeated 10 times with a set of 10 different random seeds, which remained constant among various experiments. This way, we could fulfill the above assumptions by ensuring that the only difference between similar experiments lies in the varying parameter values. Specifically, the initial weights and the order of serving the training examples were only dependent on the random seed when running on the same hardware setup.

Consistently, the mean and standard deviation of the output metric values of the 10 statistical runs were presented on line plots in the form of a solid line (mean) with a shaded area around it ($\pm$ standard deviation range). Also, each time a particular value of the output metric appears in a table or in text, it should always, unless oppositely stated, be considered a mean value from the 10 repeated runs.

A. Software and hardware setup

We prepared the models implementation in Python programming language with the use of *PyTorch* deep learning library. The plots were obtained by the use of *matplotlib.plotly* Python package. Common architecture models (EfficientNets) were obtained by *timm* Python package [25]. Adaptive Gradient Clipping (AGC), though not eventually used to generate the presented results, was used in some initial experiments with the help of *nfnets-pytorch* Python package [2]. The complete code of the implementation of the proposed method as well as of the experimental procedure is available on github: https://github.com/philvec/grad-overdrive-pyTorch on the MIT license.

The experiments were run with a cross-platform Python 3.8 interpreter on several different physical and virtual machines. However, we always used a single GPU, multi-core CPU, and a fast SSD disk.

B. Datasets used

Full CIFAR10. The full CIFAR10 dataset [17] from https://www.cs.toronto.edu/~kriz/cifar.html consists of 60000 RGB images of size 32x32 px divided into 10 classes with 6000 images per class. There are 50000 training images and 10000 test images. When we refer to using this dataset, we mean training on the original training set and validating on the original test set.

Reduced CIFAR10. The reduced CIFAR10 dataset is a subset of the full CIFAR10 dataset, i.e., we only use 10000 training samples (first original batch) for training and the complete original validation set for validation.

CIFAR100. Analogous to CIFAR10, the full CIFAR100 dataset [17] from https://www.cs.toronto.edu/~kriz/cifar.html consists of 60000 RGB images of size 32x32 px divided into 100 classes with 600 images per class. There are 5000 training images and 1000 test images. When we refer to using this dataset, we mean training on the original training set and validating on the original test set.

VI. Experiments

A. Experiment 1: Exploration of the influence of the network activation function and the method's hyperparameters on its properties.

The particular goal of this experiment was to find initial key relationships between the network activation function, the proposed method's hyperparameters, and its early-epoch training performance. Specifically, the experiment was focused on determining a subset of commonly used CNN activation functions and possibly presenting new ones that the method works well. It was also oriented at choosing rough Gradient Overdrive automatic-adjustment algorithm hyperparameters, which should form a base for further experiments.

Setup. Constant hyperparameter values:

- neural network architecture of 3 convolutional, 3 pooling and 2 linear layers
- AdamW learning rate optimizer with learning rate parameter 0.001 (remaining parameters as in [19])
- batch SGD training algorithm with CrossEntropy Loss function, batch size 128 and number of epochs 15
- reduced CIFAR10 dataset.

Adjusted hyperparameter values:

- activation function type - one of: {*Sigmoid, TanH, ReLU, LeakyReLU, GELU, ELU, SELU, SiLU, Hardswish, Sintoix, ReSintoix, LeakySintoix, ShiftHalfSintoix*} (together 13 various activation functions were tested,)
- GO usage flag - one of {*True, False*} - if *False* (i.e. control case scenario), the following GO parameters are not applicable
- GO parameter Pm (*max probability*) - within values: {0.8, 0.9, 1.0}
- GO parameter Pc (*probability cutout*) - within values: {0.08, 0.1, 0.12, 0.14}
- GO parameter e (*variance index*) - within values: {1.5, 2.0, 2.5, 3.0}

The variable hyperparameters search was organised in a traditional grid as applicable, yielding the total number of different scenarios: $13 * (3 * 4 * 4 + 1) = 637$

Results. As the dataset is balanced and as the goal was to evaluate both **time-efficiency** of training and **final classification score**, the metrics used in this experiment include:

1) an Average of all epochs of Validation Accuracy (accuracy measured on the validation set) per-epoch-Mean value (referred to as AvgVAM),

2) a Maximum of all epochs of Validation Accuracy (accuracy measured on the validation set) per-epoch-Mean value (referred to as MaxVAM).

These metrics can be evaluated for every single experimental scenario (a set of GO hyperparameters + activation function).

Figure 3 presents average metric values for all activation functions, for each GO hyperparameter set, precisely: Figure 3a for AvgVAM and Figure 3b for MeanVAM. The results identify the sensitivity of the method's hyperparameters on activation function changes. They also determine the best GO hyperparameters in a general scenario.

Furthermore, we present best results **for each activation function**, precisely: Table I for GO hyperparameter set resulting in best AvgVAM and Table II for GO hyperparameter set resulting in best MaxVAM. The visualisations serve identifying a subset of activation functions that work best with GO method.

Comments. We draw the following conclusions from this experiment's results:

- Sigmoid activation function works best with the GO method; however, it does not apply to deep convolutional networks due to the vanishing gradient problem. We observe its poor absolute results also in our training scenario. A classic ReLU activation turns out to work the best along all common functions used in modern networks. We choose 4 o the best results to be used in the following experiment.
- Several patterns can be observed in terms of GO hyperparameter search. $Pm = 1.0$ (right) is the best bet in the general case, values of $Pm = 0.8$ (left) are the best yet solitary, suggesting specificity.
- The same set of hyperparameters always implies both significant speedup and higher score, as e.g., different GO hyperparameters were chosen for ReLU, depending on the metrics used.
- Though some sets of GO hyperparameters can be discarded due to poor performance, but our method generally experiences high sensitivity to those hyperparameter changes. After using best values from table 3b as a starting point, we suggest performing a small search on the particular architecture and dataset.

B. Experiment 2: Comparison of the qualitative results on a larger network versus a small network from the previous step.

The particular goal of this experiment was to investigate the ability of the proposed method to work well on CNN architectures of different scales. We also observe the hyperparameters' sensitivity on that architecture and scale changes. The experiment compares the performance of GO from the previous step to its performance on a much more sophisticated yet classic ResNet architecture while using the best subset of the same hyperparameters. The activations marked in previous Experiment as interesting (**bold** in Tables I and II) are being tested along with its best GO hyperparameters versus a larger residual network.

Setup. Constant hyperparameters values:

- neural network architecture of 20 layer ResNet (as in [10] sec.4.2. *CIFAR-10 and Analysis*) with variable activation function parameter
- AdamW learning rate optimizer with constant learning rate parameter 0.001
- batch SGD training algorithm with CrossEntropy Loss function, batch size 128 and number of epochs 30
- full CIFAR10 dataset (see section V-B for details).

Adjustable hyperparameters values:

- GO usage flag - one of {*True, False*} - if *False* (i.e. control case scenario), the following GO parameters are not applicable
- activation function and GO parameters $(activation, Pm, Pc, e)$ - one of the consistent sets:
  - (*ReLU*, 0.9, 0.12, 3.0)
  - (*LeakySintoix*, 0.8, 0.1, 3.0)
  - (*ShiftHalfSintoix*, 0.9, 0.08, 2.0)
  - (*LeakyReLU*, 1.0, 0.08, 1.5)

The adjusted hyperparameter were organised in a traditional grid as applicable, yielding the total number of different scenarios: $2 * 4 = 8$

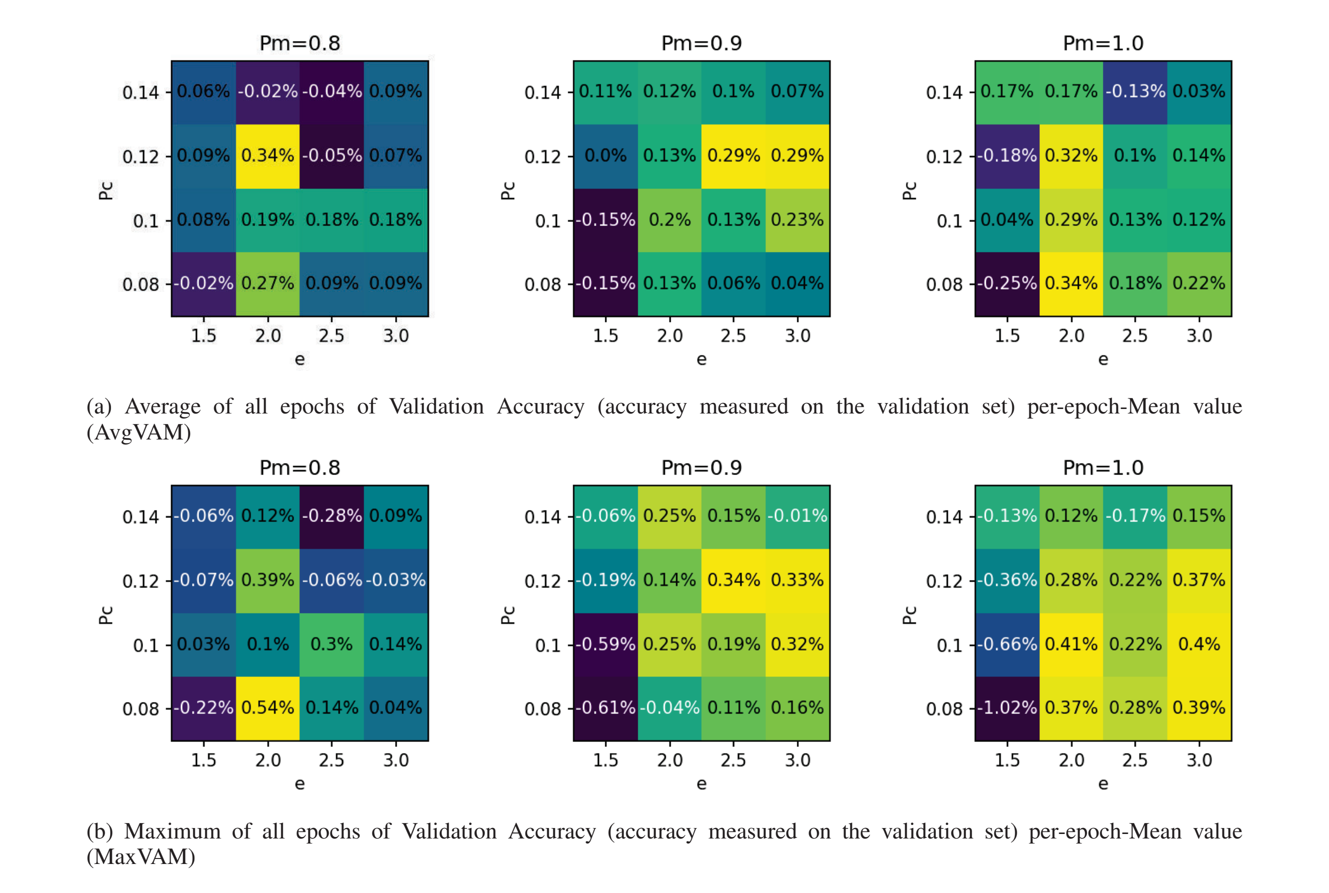
Results. In this experiment, we use the same metrics as before for comparing the results with Experiment 1.

Tables III and IV present metric values (AvgVAM and MaxVAM respectively) for each of the selected activation function and the corresponding hyperparameter set. They compare the GO method's ability to speed up the training of larger networks to the one observed in Experiment 1. We aim at determining if GO hyperparameters depend on the model scale.

Furthermore, we plot Validation Accuracy (accuracy measured on the validation set) Mean (VAM) after each epoch in Figure 4, to observe how the difference between GO and no-GO runs changes through training.

Comments. Below we list the conclusions from this experiment.

- We observe a significant speedup of GO vs. reference cases when using LeakyReLU (a) and ShiftHalfSontoix (d). Other activation functions experience only a slight boost in the first epochs of training but, unfortunately, lose their pace later on.
- We did not observe a significant increase in GO's eventual best score (MaxVAM).
- We observe a slight yet significant correlation between the results of this experiment and the results of Experiment 1. The best hyperparameters from Experiment 1 also increase the AvgVAM score when applied to a larger network. It suggests that though the method should be further improved, it can probably be used with commonly used activation functions.
- Two of our custom activation functions experience a boost in the first epochs of training, increasing the AvgVAM metric. These results are promising yet unsatisfactory and suggest further improvements have to be applied to our GO method.



(a) Average of all epochs of Validation Accuracy (accuracy measured on the validation set) per-epoch-Mean value (AvgVAM)

(b) Maximum of all epochs of Validation Accuracy (accuracy measured on the validation set) per-epoch-Mean value (MaxVAM)

Fig. 3: Relative average metric values for reduced CIFAR10 dataset for all activation functions, for each GO hyperparameter set. Lighter squares represent higher values. Several patterns can be observed, including: $Pm = 1.0$ (right) is the best bet in the general case, values of $Pm = 0.8$ (left) are the best yet solitary, which suggest specificity.

C. Experiment 3: Evaluation of the method qualitative and quantitative results on a large architecture with modern regularisation techniques.

The particular goal of this experiment was to determine if the method is well suited also for much larger, modern CNN architectures and complex image classification problems. It thus aimed at deciding if the hyperparameters tuned in a much more trivial setup of the previous experiment can be used to solve modern problems with different CNN architectures. Unlike in previous experiments, we used only three different seeds this time due to the considerable computational cost of each run. However, it should not violate the assumption about statistical reliability, as the training with larger networks is believed to be more stable and easier to reproduce. Indeed, the results of the consecutive runs are very similar as the shaded areas of the standard deviation range are very narrow in each case. **Setup**. Constant hyperparameters values:

- CIFAR100 dataset (see section V-B for details). The images were upsampled to $224 \times 224px$ resolution and treated with data augmentation: random horizontal flipping, random cropping (with 28px padding on each side),
- AdamW learning rate optimizer with constant learning rate parameter 0.001,
- neural network architecture: EfficientNet_B0 [24] with SiLU activation function and batch normalisation},
- batch SGD training algorithm with CrossEntropy Loss function, batch size 256 and number of epochs 40,
- GO hyperparameters - the same as for best results for SiLU activation function from Experiment 1, precisely: $Pm = 0.8, Pc = 0.14, e = 1.5$.

Adjusted hyperparameter values:

- GO usage flag - one of $\{$*True, False*$\}$ - if *False* (i.e. control case scenario), the GO parameters are not applicable.

Results. Both validation and training mean accuracy plot after each epoch of training is presented in Figure 5. It serves to observe how the difference between GO and no-GO runs changes through training.

Comments. Below we list the conclusions from this experiment.

- we observed a slight yet constant advantage of GO in the training curves.
- A slight advantage also in validation curves can be observed in the very first 5 epochs - this is a promising result as it shows that GO's positive impact on training efficiency is not reduced when used with large modern DNNs.

activation function	**AvgVAM without GO**	**max AvgVAM with GO**	**percentage increase**	**best GO hyperparams**		
				Pm	**Pc**	**e**
Sigmoid	0.24702	0.26419	$+6.952$ %	1.0	0.08	1.5
ReLU	**0.45826**	**0.47466**	**$+3.578$ %**	**0.9**	**0.12**	**3.0**
LeakyReLU	**0.47044**	**0.47588**	**$+1.156$ %**	**1.0**	**0.08**	**1.5**
ShiftHalfSintoix	**0.47718**	**0.48231**	**$+1.075$ %**	**0.9**	**0.08**	**2.0**
LeakySintoix	**0.48408**	**0.48926**	**$+1.069$ %**	**0.8**	**0.1**	**3.0**
Hardswish	0.46432	0.46779	$+0.748$ %	1.0	0.1	1.5
SiLU	0.46949	0.47180	$+0.491$ %	0.9	0.14	1.5
GELU	0.46914	0.47113	$+0.422$ %	1.0	0.1	2.0
Tanh	0.50665	0.50841	$+0.346$ %	1.0	0.1	2.5
ELU	0.51234	0.51405	$+0.334$ %	1.0	0.12	2.5
ReSintoix	0.52645	0.52773	$+0.241$ %	0.8	0.08	3.5
SELU	0.53196	0.53162	-0.063 %	0.8	0.14	3.5
Sintoix	0.54254	0.54178	-0.139 %	0.8	0.14	3.0

TABLE I: Best hyperparameter sets for each activation function in terms of AvgVAM results. Activations in **bold** are the most promising, as they achieve both high absolute score and high increase ratio vs. reference case.

activation function	**MaxVAM without GO**	**max MaxVAM with GO**	**percentage increase**	**best GO hyperparams**		
				Pm	**Pc**	**e**
Sigmoid	0.32795	0.36135	$+10.18$ %	1.0	0.08	1.5
ReLU	**0.52386**	**0.55000**	**$+4.988$ %**	**0.8**	**0.08**	**2.0**
LeakyReLU	**0.53782**	**0.54506**	**$+1.346$ %**	**0.9**	**0.1**	**2.0**
Hardswish	**0.52142**	**0.52792**	**$+1.246$ %**	**1.0**	**0.12**	**1.5**
GELU	0.52358	0.52784	$+0.813$ %	1.0	0.08	2.5
Tanh	0.56931	0.57263	$+0.584$ %	1.0	0.1	2.5
SiLU	0.52417	0.52713	$+0.563$ %	0.8	0.14	1.5
ShiftHalfSintoix	0.54638	0.54936	$+0.546$ %	1.0	0.1	2.0
ReSintoix	0.57396	0.57709	$+0.544$ %	0.8	0.14	3.0
ELU	0.57262	0.57535	$+0.475$ %	1.0	0.1	2.5
LeakySintoix	0.55541	0.55709	$+0.301$ %	1.0	0.08	3.0
SELU	0.58618	0.58655	$+0.063$ %	0.8	0.14	3.0
Sintoix	0.56310	0.56316	$+0.010$ %	0.8	0.12	2.5

TABLE II: Best hyperparameter sets for each activation function in terms of MaxVAM results. Activations in **bold** are the most promising, as they achieve both high absolute score and high increase ratio vs. reference case.

activation function	**AvgVAM without GO**	**AvgVAM with GO**	**percentage increase**	**GO hyperparams**		
				Pm	**Pc**	**e**
LeakyReLU	0.79459	0.79847	$+0.487$ %	1.0	0.08	1.5
ShiftHalfSintoix	0.80814	0.81146	$+0.410$ %	0.9	0.08	2.0
LeakySintoix	0.81232	0.81242	$+0.012$ %	0.8	0.1	3.0
ReLU	0.79842	0.79579	-0.329 %	0.9	0.12	3.0

TABLE III: AvgVAM scores for each activation function, with and without GO. Compare with Table I.

- However, no significant validation set accuracy increase can be observed in the overall training process, which shows that our GO method should be further improved to meet the needs of large DNN architectures with modern regularisation techniques.

VII. Conclusions

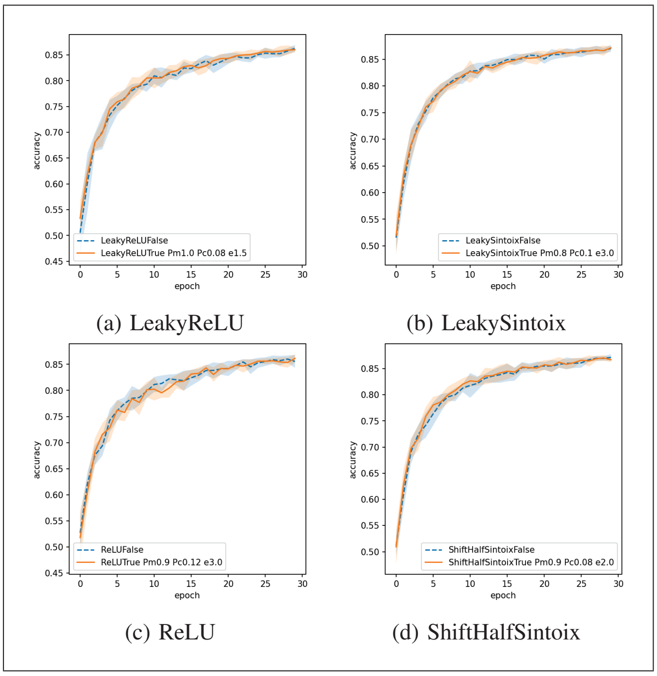
Looking back at the initial goals of the work, we can summarize what can be done to improve the results and what directions of further study seem to be promising.

The experiments presented in section VI were conducted with all the presumptions initially made. The overall experimental results are promising yet unsatisfactory to the end. We proved that our GO method could significantly increase the training speed in the initial epochs of training in simple scenarios, including ResNet networks. However, the goal of improving the training of large networks with modern regularisation and optimization techniques was not met. We seek causes of unsatisfactory results in the algorithm for adjusting the probability parameter of the GO method during training (see section III-C for details). Our method seems to work well on simple architectures but does not seem to scale well into larger ones.

For this reason, we think more focus should be taken in terms of the hyperparameters whose importance was marginalized in our work due to limited resources. These hyperparameters include: β parameter of the exponential running average used in estimating EMVar, batch size, and learning rate. As recent attempts to speed up the training of large networks into image classification problems require extensive computations and specialized hardware (including multiple GPUs or TPU cores), we think these hyperparameters should be optimized

activation function	MaxVAM without GO	MaxVAM with GO	percentage increase	GO hyperparams		
				Pm	Pc	e
ReLU	0.85997	0.86128	+0.152 %	0.9	0.12	3.0
LeakySintoix	0.87037	0.87138	+0.116 %	0.8	0.1	3.0
ShiftHalfSintoix	0.87139	0.87028	−0.127 %	0.8	0.08	2.0
LeakyReLU	0.86274	0.86010	−0.305 %	1.0	0.08	1.5

TABLE IV: MaxVAM scores for each activation function, with and without GO. Compare with Table II.



(a) LeakyReLU

(b) LeakySintoix

(c) ReLU

(d) ShiftHalfSintoix

Fig. 4: Learning curves (accuracy measured on the validation set vs. epoch) for the best GO hyperparameters on 20-layer ResNet network, CIFAR10 dataset. 'False' depicts the control case without GO, 'True' illustrates the use of GO. Solid lines represent mean values; shaded areas represent the ± standard deviation range. We observe a slight speedup of GO vs. reference case when using LeakyReLU (a) and ShiftHalfSontoix (d).

in an environment of richer resources.

The experiments presented in section VI were conducted with all the presumptions initially made. The overall experimental results are promising yet unsatisfactory to the end. We proved that our GO method could significantly increase the training speed in the initial epochs of training in simple scenarios, including ResNet networks. However, we did not achieve the goal of improving the training of large networks with modern regularisation and optimization techniques. We seek causes of unsatisfactory results in the algorithm for adjusting the probability parameter of the GO method during training (see section III-C for details). Our method seems to work well on simple architectures but does not seem to scale well into larger ones.

Although the solution is not entirely successful, we do not consider the topic to be exhausted. As we seek causes of unsatisfactory results in our GO automatic adjustment method, further attempts should be made to improve it. As the general idea of overdriving the gradients to reduce the phenomenon of scrabbling weights still seems promising, different approaches to its adjustment can be tested.

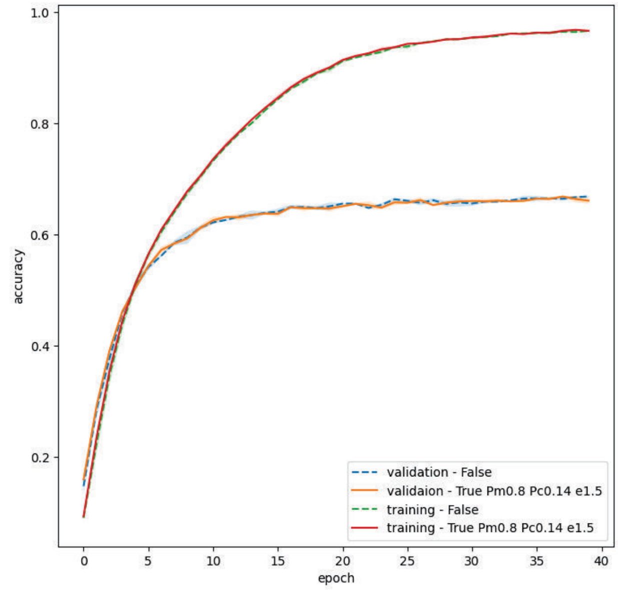


Fig. 5: Learning curves (accuracy measured respectively on training and validation sets vs. epoch) on EfficientNet_B0 baseline network, CIFAR100 dataset. 'False' depicts the control case without GO, 'True' illustrates the use of GO. Solid lines represent mean values; shaded areas represent ± standard deviation range. A slight yet constant advantage of GO can be observed in training curves (solid red and dashed green).

Furthermore, the use of TPUs (Tensor Processor Units) would enable fast training of even larger models on ImageNet ILSVRC [23]. We think of this dataset as an ultimate benchmark also for our GO method, which may eventually prove its full applicability if the spotted problems have been resolved.

Eventually, though measuring and optimizing the computational performance of our method was not in the scope of this work, such research should be considered in future work. We found a slight increase in overall computation time (up to around 10%) versus a reference case when using AdamW optimizer. We believe this load can be reduced even more by profiling and optimizing the code.

References

[1] L. Bottou, *Stochastic gradient descent tricks*, In: Montavon G, OrrGB, Mueller KR (eds) Neural Networks: Tricks of the Trade: Second Edi-tion, Springer Berlin Heidelberg, Berlin, Heidelberg, pp 421–436, 2012.

[2] V. Balloli, *A pytorch implementation of nfnets and adaptive gradient clipping.* https://github.com/vballoli/nfnets-pytorch, 2021.

[3] R. Pascanu, T. Mikolov, and Y. Bengio, *On the difficulty of training recurrentneural networks,* 2013.

[4] P. Cheridito, A. Jentzen,and F. Rossmannek *Non-convergence of stochastic gradient descent in the training of deep neural networks*, Journal of Complexity,vol 64, Elsevier BV, 2021.

[5] A. Jentzen and T. Welti, *Overall error analysis for the training of deep neural networks via stochastic gradient descent with random initialisation*, archiv, 2020.

[6] Y. Cao and Q. Gu, *Generalization Bounds of Stochastic Gradient Descent for Wide and Deep Neural Networks*, archiv, 2019.

[7] S. Ioffe and C. Szegedy, *Batch normalization: Accelerating deep network training by reducing internal covariate shift,* 2015.

[8] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov,*Improving neural networks by preventing co-adaptation of feature detectors,* 2012.

[9] P. Jain, S. M. Kakade, R. Kidambi, P. Netrapalli and A. Sidford,*Accelerating Stochastic Gradient Descent For Least Squares Regression*, archiv, 2018.

[10] K. He, et al., “Deep residual learning for image recogni-tion,” 2015

[11] S. Hooker, *The Hardware Lottery*, Archiv, 2020.

[12] Z. Zhu, J. Wu, B. Yu and L. Wu and J. Ma, *The Anisotropic Noise in Stochastic Gradient Descent: Its Behavior of Escaping from Sharp Minima and Regularization Effects*, archiv, 2019.

[13] , P. Zhao and T. Zhang, *Accelerating Minibatch Stochastic Gradient Descent using Stratified Sampling*, archiv,2014.

[14] R. Johnson and T. Zhang, *Accelerating Stochastic Gradient Descent using Predictive Variance Reduction*, Advances in Neural Information Processing Systems, vol. 26 C. J. C. Burges and L. Bottou and M. Welling and Z. Ghahramani and K. Q. Weinberger, Curran Associates, Inc., 2013.

[15] M. Schmidt. N. Le Roux and F. Bach, *Minimizing Finite Sums with the Stochastic Average Gradient*, archiv 1309.2388, 2013.

[16] A. Defazio, *A Simple Practical Accelerated Method for Finite Sums*, archiv 1602.02442, 2016.

[17] A. Krizhevsky, *Learning multiple layers of features from tiny images,* University of Toronto, 2012.

[18] A, Defazio and L. Bottou, *On the Ineffectiveness of Variance Reduced Optimization for Deep Learning*, archiv 1812.04529, 2019.

[19] D. P. Kingma and J. Ba *Adam: A Method for Stochastic Optimization*, archiv 1412.6980, 2017.

[20] I. Loshchilov and F. Hutter *Decoupled Weight Decay Regularization*, archive 1711.05101,2020.

[21] R. M. Schmidt, F. Schneider and P. Hennig *Descending through a Crowded Valley – Benchmarking Deep Learning Optimizers*, archiv 2007.01547, 2020.

[22] D. Musso, *Stochastic gradient descent with random learning rate*, archive 2003.06926, 2020.

[23] O. Russakovsky, et al. *ImageNetLarge Scale Visual Recognition Challenge,* International Journal of Computer Vision (IJCV), vol. 115, no. 3, 2015.

[24] M. Tan and Q. V. Le, *Efficientnet: Rethinking model scaling for convolutionalneural networks,* 2020.

[25] R. Wightman, *Pytorch image models,* https://github.com/rwightman/pytorch-image-models, 2019.

A Fuzzy Radiotherapy Planning Dose Distribution Model Using Interior Point Methods

Aurelio R. L. Oliveira, Jackeline del Carmen Huaccha Neyra

***Abstract*—** In radiotherapy planning one of the goals consists in minimize the radiation dose delivered in the patient, in particular in the healthy tissues and organs. On the other hand, it is necessary to deliver as much radiation as possible in the tumor. In order to achieve these goals we use a linear programming model to formulate this problem and the intensity modulated radiation therapy technique which allows the delivery of nonuniform radiation flow. In practice, the amount of dose to be delivered varies according with the specialist. Therefore in this work, the doses are considered as triangular fuzzy numbers. Besides, the surprise function, that can be view as a penalty for the constraints violation, models these fuzzy constraints in a non-linear function that must be minimized. As a result, a mathematical model with non-linear and convex objective function with bounding constraints for the radiation flow is obtained. We develop a specially tailored interior point method for this mathematical model. It was implemented in MATLAB and the numerical experiments are performed in real world large-scale problems. The numerical results show that the developed method provides favorable solutions for the dose distribution problem in comparison with previous approaches.

***Keywords*—** Radiation therapy planning, Fuzzy numbers, Interior point methods, Surprise function.

Aurelio Oliveira is with the University of Campinas, Brazil (e-mail: aurelio@ime.unicamp.br).

Special and General Relativity and the Relativistic Ether as Observer Dependent by the Retardation

Azzam Almosallami

Zurich, Switzerland

a.almosalllami71@gmail.com

Abstract

In this paper I'll show how the relativistic effect in SRT must be observer dependent which is leading to field and retardation, and that is leading to the wave-particle duality and the uncertainty principle by the vacuum fluctuation. In this I propose a new transformation by translating the retardation according to the invariance by the entanglement which is leading to the relativistic ether from the point of view of the quantum vacuum which is leading to the wave-particle duality and the uncertainty principle by the vacuum fluctuation. According to my transformation, there two pictures for the moving train, and these two pictures are separate in space and time as a result of the retardation but they are entangled by the invariance of the energy momentum. That will lead also to explain the double slit experiment from the point of view of quantum theory. In my new transformation, I propose there is no space-time continuum, as in special relativity; it is only time, and space is invariant. That leads to the new transformation being vacuum energy dependent instead of relative velocity dependent as in Einstein's interpretation of the Lorentz transformation equations of the theory of special relativity. Furthermore, the Lorentz factor in my transformation is equivalent to the refractive index in optics. That leads to the disappearance of all the paradoxes of the theory of special relativity: The Twin paradox, Ehrenfest paradox, the Ladder paradox, and Bell's spaceship paradox. Furthermore, according to my interpretation, one could explain the experimental results of quantum tunneling and entanglement (spooky action), Casimir effect, and Hartman effect. Also according to that by my equivalence principle, dark matter and dark energy are explained, and no need to propose dark matter and dark energy, and as a consequence of that, the cosmological constant problem will be solved.

Kaywords: Special relativity, General relativity, relativistic ether, retardation, vacuum fluctuation, wave-particle duality, uncertainty principle, Entanglement.

Theory

In my paper [1], I have reached to my new=transformation−

$$= \quad '-'$$

$$=$$

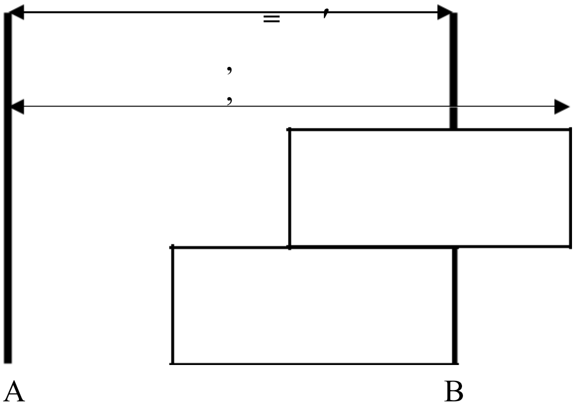


Figure 1: illustrates the location of the moving train during the motion for the observer stationary on the ground according to his time ground clock reading t, and the location of the moving train for the observer stationary on the moving train according to his time clock reading t' stationary on moving train.

My transformation expresses about the clock retardation according to the invariance of the energy momentum by the entanglement. That will lead to the wave-particle duality and the uncertainty principle by the vacuum fluctuation [1-6].

In my transformation space is invariant, and that means there is no length contraction as proposed by Lorentz transformation in order to keep on the reality is observer independent according to the Minkowski space-time. As Minkowski said about his work in relativity, ***"The views of space and time which I wish to lay before you have sprung from the soil of experimental physics, and therein lies their strength. They are radical. Henceforth space by itself, and time by itself, are doomed to fade away into mere shadows, and only a kind of union of the two will preserve an independent reality*** [18]**."** And we find that will lead to a great problem in physics specially when we are trying to reach to the unified theory.

According to my transformation we get the relativistic ether from the point of view of quantum vacuum which is quantum field theory, and this relativistic ether is observer dependent. Engelhardt [9] obtained the dispersion relation according to Maxwell's ether theory of light by Galilean transformation according the phase velocity. What I found in my paper, if we translated the retardation according to the invariance by the entanglement by considering space is invariant, in this case we get a vacuum fluctuation as a result of the wave-particle duality and the uncertainty, and according to that we get the relativistic ether as observer dependent from the point of view of the quantum vacuum and the vacuum fluctuation. In this case we get the speed of light must be constant in the local classical vacuum, and globally the speed of light is not constant by fluctuate. And that will lead to Doppler effect for photons is the consequence of the energy and momentum exchange between the atom and the photon: a central role is played by **the quantum energy jump ΔE of the transition (a relativistic invariant) same as showed by** Schrodinger in 1922 [7], and the same we can explain the aberration.

Now according to my transformation when the moving train at constant speed v arrives pylon B as in Fig. (1) for the observer stationary on the moving train, in this case we get for the observer and stationary on the moving train

$'$

For the observer stationary on the ground,= at this momentum, the front of the moving train passed pylon B and the moving train is at distance from pylon A

$$= \quad = \quad ' \qquad (4)$$

In this case we get that the term $2'$ will be equal to zero. Here in my transformation x and x' represent the light path for the two observers, the observer stationary on the ground and the observer stationary on the moving train.

Here L is the length of the moving train which is invariant for the both the two observers on the ground and the observer stationary on the moving. In my theory space is invariant. That means both the two observers will agree at the length of the moving train to be L during the motion same as if the train is stationary.

In this case when an object inside the moving train leaves the boundaries of his moving train to the ground, this case there must be a vacuum fluctuation by the wave-particle duality and the uncertainty as a result of the retardation, and thus in this case the clock of the object is reading time t' as a result of the retardation, and that will be explained in more details in my equivalence principle. In this case we get

$$' = \quad -$$

and in this case when the object leaves the boundaries of the moving train to the ground, in this case there must be a vacuum fluctuation, and when we make a localization at this moment, the object will be at distance on the ground from pylon A

$'$

not at a distance because at this moment, the object left the boundaries space of the space of the ground. My transformation illustrates the theory of Feynman and

moving train to the ′

how time is moving = forward not backward. In this case for the object for itself, the object is

transformation from a distance on the ground at to a distance in a zero-time

$' \;=$ $'=$

separation.

Now by considering the length of the moving train is invariant for both the two observers stationary on the ground and stationary on the moving train, in this case we get for the observer stationary on the moving train, the speed of light is c locally where in this case we have according to his clock locally

$$- \quad - \quad -$$

According to that we get according to my transformation that ***"the light speed is constant in the local classical vacuum".***

For the observer stationary on the ground there are two velocities are measured for the light beam globally when it leaves the space of the moving train, the phase and group, which are measured globally as a result of the retardation as

And the group velocity′ as a result of the=vacuum fluctuation, and in this case the uncertainty principle pays the rule as– a result– of the retardation,– in this case we get

That's how the speed of light globally according to my transformation is not constant by fluctuate!

THE RELATIVISTIC ENERGY-MOMENTUM

The relativistic kinetic energy of the moving train relative to an observer stationary on ground is given according to the equation

$$\mathbf{E} \quad = \sqrt{}\,^{\mathbf{E}}\,\mathbf{v} - \mathbf{E}\,,$$
$$- \quad \mathbf{e}$$

E = **m c** and **E** is the relativistic kinetic energy. Now by solving this equation in terms of v, we get

,

$$E + \quad E\ E$$

$$v = \sqrt{E} \quad \overline{\quad + E \quad} \quad c$$

When E ≪ E we get the classi~~cal kinetic en~~ergy E = m v

According to Eq. (11) there is no way for the moving train to reach to the speed of light locally on ground.

The relativistic momentum of the moving train relative to the observer stationary on the ground is given according to the equation $\mathbf{mv}$ $\mathbf{P} = \sqrt{\ } - \mathbf{v}\cancel{\mathbf{e}}$

Now substitute the value of v from Eq. (11) in Eq.(12), we get

$$\frac{\mathbf{P\ c}}{\quad} = \mathbf{E}$$

Now when $\mathbf{m} =$ we get $E+$ $\mathbf{m\ c}$ $\underline{\mathbf{hv}}$

$$\mathbf{P} = \quad \mathbf{c}$$

Now if we substitute from Eq. (10) the value of **E** $=$ ~~E0~~ $\underline{\cancel{\mathbf{v2}}}$ − **E** in Eq. (13), we get

And from that we get $\mathbf{E = P\ c + m\ c}$ $\sqrt{\ } - \mathbf{c\ 2}$

Where $\mathbf{P\ c = E - m\ c}$

$$\mathbf{E =} \quad \mathbf{E}$$

Now we have here the relativistic momentum of the moving train relative to the observer

stationary on the ground is $\sqrt{\ } \quad - \quad \cancel{\mathbf{e}}$ $\frac{\underline{\mathbf{m}}\ \mathbf{v}}{\sqrt{\ } - \mathbf{v}\cancel{\mathbf{e}}}$

which leading to Eq. (13). While the classical momentum of the moving train for itself according to the definition of the proper time in special**P** relativity= which**mv** is leading to the retardation in my transformation is given as where the momentum of the moving train for itself is the classical momentum, **m** is the rest mass of the moving train. Here we proposed the self-momentum of the moving train in order to get the retardation which is proven experimentally which is translated in my transformation.

Now by the reciprocity we find the momentum of the observer stationary on the ground relative to the observer stationary on the moving train

$$\mathbf{P} \quad - \quad \mathbf{y} \quad \mathbf{y} = \mathbf{m\ v}$$

which is the classical momentum by the reciprocity, where here m is the rest mass of the observer stationary on the ground. That illustrates the relationship between my transformation and the Galilean transformation, where in this case my transformation will lead to the Copenhagen school by translating the retardation according to the invariance by the entanglement which is leading to the wave-particle duality and the uncertainty principle by the vacuum fluctuation. That explains how *in classical mechanics, a special status is assigned to time in the sense that it is treated as a classical background parameter, external to the system itself. This special role is seen in the standard formulation of quantum mechanics. It is regarded as part of an a priori given classical background with a well defined value. In fact, the classical treatment of time is deeply intertwined with the Copenhagen interpretation of quantum mechanics, and, thus, with the conceptual foundations of quantum theory: all measurements of observables are made at certain instants of time and probabilities are only assigned to such measurements* [19].

According to that we could solve the problem of time [19] according to my transformation. Where in my transformation, the Lorentz factor by translating the retardation≤γ-≤according to invariance by the entanglement is leading to the probability in QM where .

THE EQUIVALENCE PRINCIPLE AND THE RELATIVISTIC ESCAPE VELOCITY

According to my transformation by translating the retardation according to the invariance by the entanglement, we find that the moving train is accelerated or decelerated according to the radiation exchange. That's how *in quantum mechanics, the analogue of Newton's law is Schrödinger's equation for a quantum system (usually atoms, molecules, and subatomic particles whether free, bound, or localized). It is not an algebraic equation, but in general a linear partial differential equation, describing the time-evolution of the system's wave function (also called a "state FUNCTION").* According to that we can reach to real equivalence principle as the in the following.

Suppose now the train is at rest, and after that the velocity of the moving train is changed from zero to v to move with kinetic energy locally on the ground relative to an observer stationary on the ground. According to my equivalence principle as a result changing the velocity of themovingtrainfrom0tovwithkineticenergy,thevelocityofstationaryriderinsidemoving train must be changed also from 0 to v locally on the moving train as a result of inertia. Let's see how that will happen according to my equivalence principle. According to my transformation as a result of retardation we get when the clock inside the moving train reads locally inside the moving train for the observer stationary on the moving train, at this momentum for the observer stationary on the ground his clock reads t where in this case we get from Eq. (4) according to the retardation -

This equation illustrates that the moving train is delayed for the observer stationary inside the moving train in space and time on ground comparing to where the train is now on the ground for the observer stationary on the ground according to his space and time as illustrated in Fig. (1). According to my transformation, there are two pictures for the moving train, and these two pictures are separated in space and time but entangled with each other's by the invariance of the energy momentum.

From Eq. (4) for the light beam inside the moving train relative to an observer stationary on the ground we get the frequency of the light beam must be changed as a result of the retardation comparing to the same light beam if it was transmitted on the classical space of the ground relative to the earth observer, where in this case we get for the observer on ground, the measured frequency of the light beam as a result of the retardation = will be

–

And from that the energy of the photon will be different relative to the observer stationary on the ground as a result of the retardation, where

And from that we get ′ – $h\,' = - \quad h$

From that as a result of the wave=-particle duality, we can compare the rest mass energy of the stationary observer inside the moving train relative to reference frame of the observer stationary on the ground to get

where from that we get ′= –

′ –

This difference of energy must Δ represent=– the=relativistic– kinetic energy of the rest rider inside the moving train as a result of inertia let the velocity of the rider changes from 0 to v locally on the moving train as a result the velocity of the moving train changes from 0 to v locally on the ground.

Now substitute in Eq. (11) according to the invariance of the energy momentum and

instead ofwe have now from Eq. (16) ′= –then we get = Δ

$$' = \sqrt{\frac{\Delta \quad + \quad \Delta \; '}{\Delta \; +'}} \; \mathbf{c}$$

Now substitute from Eq. (17) **ΔE = E − E** ′ in Eq. (18), we get

$$' = \sqrt{\frac{- \; ' \qquad + \; - \; ' \qquad '}{- \; '+ \; '}}$$

$$' \quad \sqrt{} \quad = \quad ' \quad +-' \; +' \quad =-'$$

$$= \; ' \quad =-'$$

$$= \sqrt{} \; ' \qquad '$$

Thus we get $' = \quad - \quad \sqrt{} \quad - \quad -$

Now substitute from Eq. (16)

$$\_\mathrm{v}$$

Now if we substitute in Eq. (20) $' = \sqrt{} \; - \quad -$

We get , where according to that when the velocity of the moving train changes 0 to v with relativistic′ kinetic energy locally on the ground, in this case the velocity of the stationary rider on the = moving v train will change also from 0 to v locally on the moving train according to

the invariance of the energy momentum.

Now from Eqs. (17)&(20), we can derive the relativistic escape velocity as in the following. If we consider the relativistic kinetic energy in Eq. (17) equals to the gravitational potential, where

And $\Delta \; =$

in this case we get $=$

And from that we get $= \qquad - \; -$

$$- \;=(- \;)$$

Now substitute from Eq. (21) in Eq. (20), we get the relativistic escape velocity locally of the free fall object under the gravitational field. In this case we get

$$- \quad \text{measured,}\sqrt{}$$

Now globally we have two velocities can be the first velocity is the phase velocity, where in this case the phase velocity of the free fall object globally is given as

$$=(\; - \quad) \sqrt{} \quad -$$

$$-\boldsymbol{h} \; -$$

And when we make a localization, in this case we get globally as the motion in linear dispersion and this case we get the group velocity and phase velocity are equal, and in this case the group velocity is given same as in Eq. (22). Now during the free fall, we have here a vacuum fluctuation, which is equivalent to motion in nonlinear dispersion, and in this case the uncertainty principle plays the rule, where in this case even if we start with a fairly localized “particle”, it will soon loose this localization. According to that the group velocity is not equal to the phase in case of nonlinear dispersion, and in this case the group velocity is given according to my transformation as

$$- \; - \qquad = \; - \qquad \sqrt{} \quad -$$

Now at strong gravitational field at the gravitational radius , we get locally . Now globally we get by the Lorentz factor the escape velocity must be zero, gravitational radius at But in quantum mechanics as a result of the treatment of matter in quantum mechanics as having properties of waves and particles. One interpretation of this duality involves the Heisenberg uncertainty principle, which defines a limit on how precisely the position and the momentum of a particle can be known at the same time. This implies that there are no solutions with a probability of exactly zero (or one), though a solution may approach infinity if, for example, the calculation for its position was taken as a probability of 1, the other, i.e. its speed, would have to be infinity. Hence, the probability of a given particle's existence on the opposite side of an intervening barrier is non-zero, and such particles will appear on the 'other' (a semantically difficult word in this instance) side with a relative frequency proportional to this probability.

The Precession of Mercury’s Perihelion

Kepler’s law can be defined as

The Kepler's $=\int$ $=\int$ $=$
second law is defined as

Now we can compute the relativistic Mercury precession according to my transformation and my equivalence principle as in the following.

when we make a localization according to my transformation and my equivalence principle, we get the phase velocity and the group velocity are equal globally which are given according to Eq. (22), and in this case, we get as a=result of the retardation from Eq. (6)

From Eq. (26) and from Kepler's law we get the area element by the distortion is given as

get

Thus by doing the integration by considering $=$ ^{2} we

$$dA □ □ \frac{1}{2} R^2 □ □R □ □^2 \ln \left| \frac{R □}{□} \right| □ d\theta$$

In the equation above we find that there is a singularity for the case when $=$ which is not the usual singularity at Schwarzschild radius but at gravitational radius

$/$. Now in case of weak $/$

Equation (27) represents the relativistic form in case of weak gravitational field of the element of area. The classical form is given from Eq. (24)

From my transformation from Eq. (5) according to the retardation and by my equivalence principle from Eq. (21), we get

Thus by dividing eq. (27) by $' = (-)$ we

$$' \quad ' = (+)$$

, in

$$' \qquad '$$

$/ \ll$

this case we get

$$' = (+)(+)$$

$$' = + + +$$

$$' = + +$$

And since in case of weak gravitational field $\ll$

In this case we get

$$' = (+)$$

Comparing the classical form of Kepler’s second= law from Eq. (25)

We can conclude from the relativistic form that

$$= -$$

By considering $' = (+)$

And by substituting the value of R in eq. (29) we get

$$\Delta\ '=\int + \quad - \quad \int - \quad - \quad \int$$

By doing this integration, we get

$$= \; -$$

What find the result we get in Eq. (30) is the same result derived by Gerber according to the retarded potential. The difference is that we get the same result by quantization of gravity by considering the quantization of the gravitational potential which is leading to the relativistic effect. We find also according to my equivalence principle and my transformation there is no need to propose dark matter and dark energy where they are explained as the result quantization of gravity. That is how classical physics and Newton's gravity can't explain dark matter or dark energy, and even since general relativity of Einstein is considered as classical, general relativity of Einstein can't explain dark matter and dark energy.

Sagnac effect

Sagnac effect can be explained according to my transformations by considering the t-term in my transformation.

$$= \quad ' -$$

If we considered $- =$ $\quad ' - \quad 2'$ and $+ = \quad ' + \quad 2'$, in this case we get

$$\Delta \quad = \quad (\quad)$$

$$\Delta = ' \; =$$

And since L is invariant and by

This result is exactly the same result which derived by Engelhardt [8] in explaining Sagnac effect in the framework of the ether theory.

And that explains how Schrodinger showed that the emission of a light quantum by a (flying) atom is regulated by the conservation laws of energy and linear momentum. Therefore, the Doppler effect for photons is the consequence of the energy and momentum exchange between the atom and the photon: **a central role is played by the quantum energy jump ΔE of the** transition (a relativistic invariant) [7].

The Pioneer anomaly

~ . ×

Radio metric data from Pioneer 10/11 indicate an apparent anomalous, constant, acceleration acting on the spacecraft with a magnitude m/ , directed towards the Sun [11,12]. of the Pioneer anomaly, and concluded that Turyshev [13] examined the constancy and direction the data a temporally decaying anomalous acceleration with an over 10% acceleration model. Anderson, who is improvement in the residuals compared to a constant retired from NASA's Jet Propulsion Laboratory (JPL), is that study's first author. He finds, so "it's either new physics or old physics we haven't discovered yet." New physics could be a variation on Newton's laws, whereas an example of as-yet-to-be-discovered old physics would be a cloud of dark matter trapped around the sun. Now I introduce the exact solution for the Pioneer anomaly depending on my transformation and my equivalence principle. and the Hubble's law. According to my solution, there are two terms of decelerations that controls the Pioneer anomaly. The first is produced by moving the Pioneer spacecraft through the gravitational field of the Sun, and this deceleration is responsible for varying behaviour of the Pioneer anomaly in Turyshev [13]. And according to the principle of the quantum superposition in my equivalence we find the second term is depending on the Hubble's law which is equal to the Hubble's constant multiplied by the speed of light in vacuum. This solution of the Pioneer anomaly will give us the origin of the problem of dark matter and dark energy and thus the cosmological constant problem. Sonnleitner [10] showed that how a simple calculation leads to the surprising result that an excited two-level atom moving through a vacuum sees a tiny friction force of first order in v/c. So we find here a connection between what is resulted from this paper and the Pioneer anomaly as a result of quantization of gravity in my theory by the retardation.

We find from Eq. (23) for the free fall object as a result of the vacuum fluctuation, the velocity of the free fall object must be decreased as observed globally. Eq. (23) is working in case of weak and strong gravitational field. This equation can be approximated in case of weak gravitational field as

$$- \quad =(- \quad)\sqrt{\quad}$$

Which is the same equation derived from the Schwarzschild Geometry in case of weak gravitational for the free fall object, but according to the Schwarzschild geometry this equation has no any physical meaning, because in reality it is in violation with the equivalence principle of Einstein, and also it is in violation with reality is observer independent according to Minkowski Geometry of space-time.

as

According to that for low velocities comparing to the speed of light, the difference between the predicted frequency and the reference frequency as the result of the red shift is given

$$\frac{\Delta}{\quad} \quad \frac{-}{\quad}$$

observed frequency difference Δ is depending on Eq. (31)

$$=$$

Now by considering the

$$\Delta \quad - \quad -$$

my equivalence

Where according to principle

$$- \quad \text{field} \quad \surd$$

Since in weak gravitational $= \quad -$

Thus in case of weak $\ll$ gravitational field we get

Thus from Eqs. (32)&(33) we get $- \quad = \surd$

$$[\Delta \quad -\Delta \quad] \quad \surd$$

$$= -(\quad)$$

From eq. (35) we get the observed difference frequency is less than the predicted. That means there is a slight blue shift. According to the Pioneer team calculations, the observed, two-way anomalous effect by a DSN antenna can be expressed to first order in V/C as in [1]

Thus from that and from $[\Delta$ eq. (35) $-\Delta$ we get $]$ $= -[\Delta \quad -\Delta \quad]$

By DSN convention $= -$

$$\surd$$

By considering in Eq. (37) $= -$ we get $-(\quad) \quad = \quad ,$

$$\surd$$

And from that we get $-(\quad) \quad = \quad ,$

√

Which is equal to ′ = −

−

′ /

represents the√ distance between the spacecraft and the Sun, and thus we find
In Eq. (38) we find r =− / /
the deceleration of the spacecraft is depending on the distance of the spacecraft from the Sun.

Now by considering = . × ≅ . = are respectively the
computing
gravitational constant and the mass

′− =−. × − / ′− =−. ×

(1983-1990) of the whole= observation× period of Pioneer= 10, its× radial distance from the Sun changes from to . Thus by from Eq. (38), we get and

Analogous computations for Pioneer 11, as checking point, show the following. Full time of
get and
observation of Pioneer 11 is shorter so observational period is taken from 1984 to 1989, with observational data from the same source [3]. Radial distances for beginning and end of the period

are − , and . By using Eq. (38) we
We have seen that the deceleration of the pioneer 10/11 anomalies is decreased depending on the distance from the Sun as from Eq. (38), and that what is causing the varying behavior of the Pioneer anomalies according to Turyshev [7]. According to the period of observation 7.5 years
from (1983-1990) as noted by Anderson [13], we find for the Pioneer 10 is given as

′ . × − − . × − − ·−

Markwardt [14] obtained an improved fit of Pioneer 10 data when estimating a jerk of

−. × − / . which is exactly· same as in my calculations. Also Toth [15] obtained′
which is agreed with my calculations.

·− =−. × /.

Now there is another term must be added to the Pioneer anomaly in Eq. (38) according to the principle of the quantum superposition in my equivalence. This term is related to the Hubble's

law. We have from Hubble=s this deceleration is given according to the equation

Where is the deceleration is caused by the Hubble, where is this case since the spacecraft is going far away from the Sun, in this case it observed for an observer on ground, there is a slight blue-shift given according to the Eqs. (38)&(39). If the spacecraft is in a free fall toward the Sun, in this case, it will be observer a slight red-shift which is given also according to Eqs. (38)&(39). According to that we get the full Pioneer anomaly is given according to

–

/

Hubble constant, √which used a new infrared camera on the Hubble Space

An estimate of the =– – ⁄
Telescope (HST) to measure the distance and redshift for a collection of astronomical objects, gives a value of H= 73.8 ± 2.4 (km/s)/Mpc or about –=–.× – [16,17]. Thus

from Eq. (40) we get for the Pioneer 10 at distance == .· ± or⁄ /

This quantity is very agreed with the observed Pioneer 10 acceleration (at t=11 years of lunch),
in fig. (1) taken =from. Turyshev [13]=. **years of lunch, we get**
At a distance = – . × – . × = – – – –

. × _ ⁄

This quantity is agreed with the observed Pioneer 10 acceleration (at t=18 years of lunch), in fig.
(2) taken from Turyshev [13].

We find from my transformation by translating the retardation according to the invariance by the entanglement which is leading to the wave-particle duality and the uncertainty principle by the vacuum fluctuation, the gravitational field is expressed according to the energy fluctuation, the vacuum energy fluctuation effectively gives a correct explanation of dark energy and dark matter, where in this case dark matter and dark energy are explained , and that will provide a solution to the cosmological constant problem. Figure (3) illustrates the predicted Pioneer 10 anomaly according to Eq. (40).

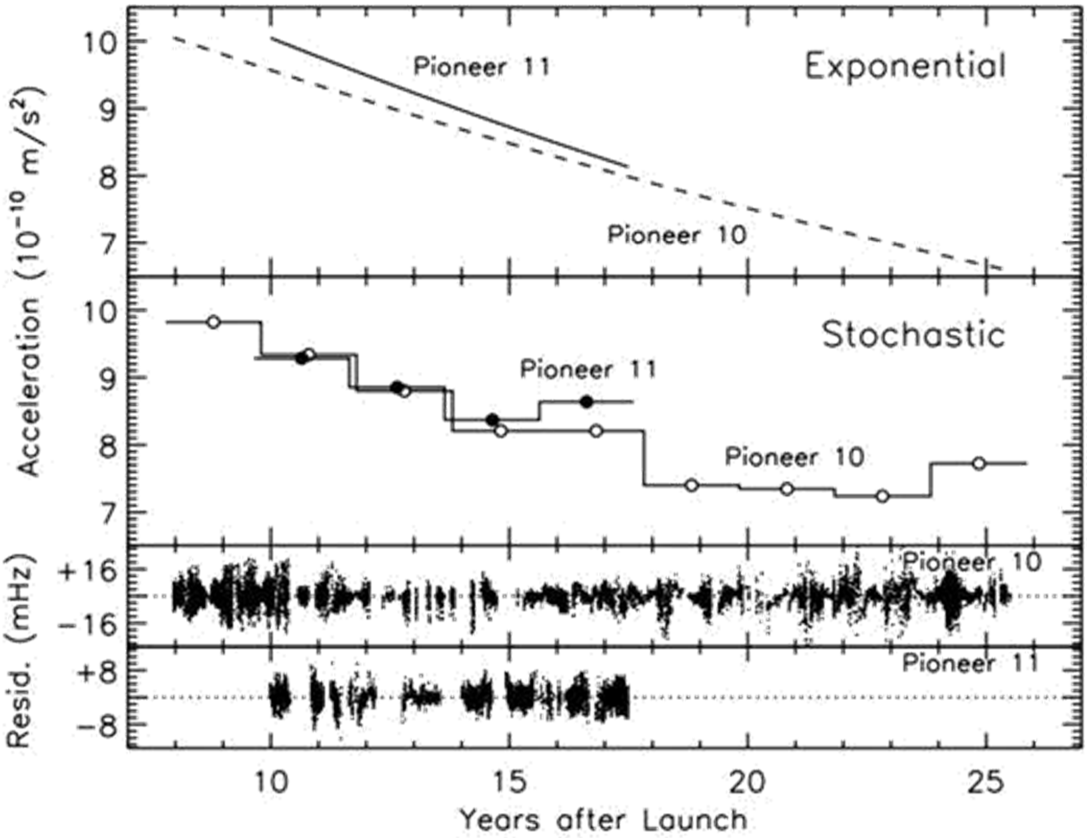


Fig. (2): Top panel: Estimates of the anomalous acceleration of Pioneer 10 (dashed line) and Pioneer 11 (solid line) using an exponential model. Second panel: Stochastic acceleration estimates for Pioneer 10 (open circles) and Pioneer 11 (filled circles), shown as step functions. Bottom two panels: Doppler residuals of the stochastic acceleration model. Note the difference in vertical scale for Pioneer 10 vs. Pioneer 11. Turyshev [13].

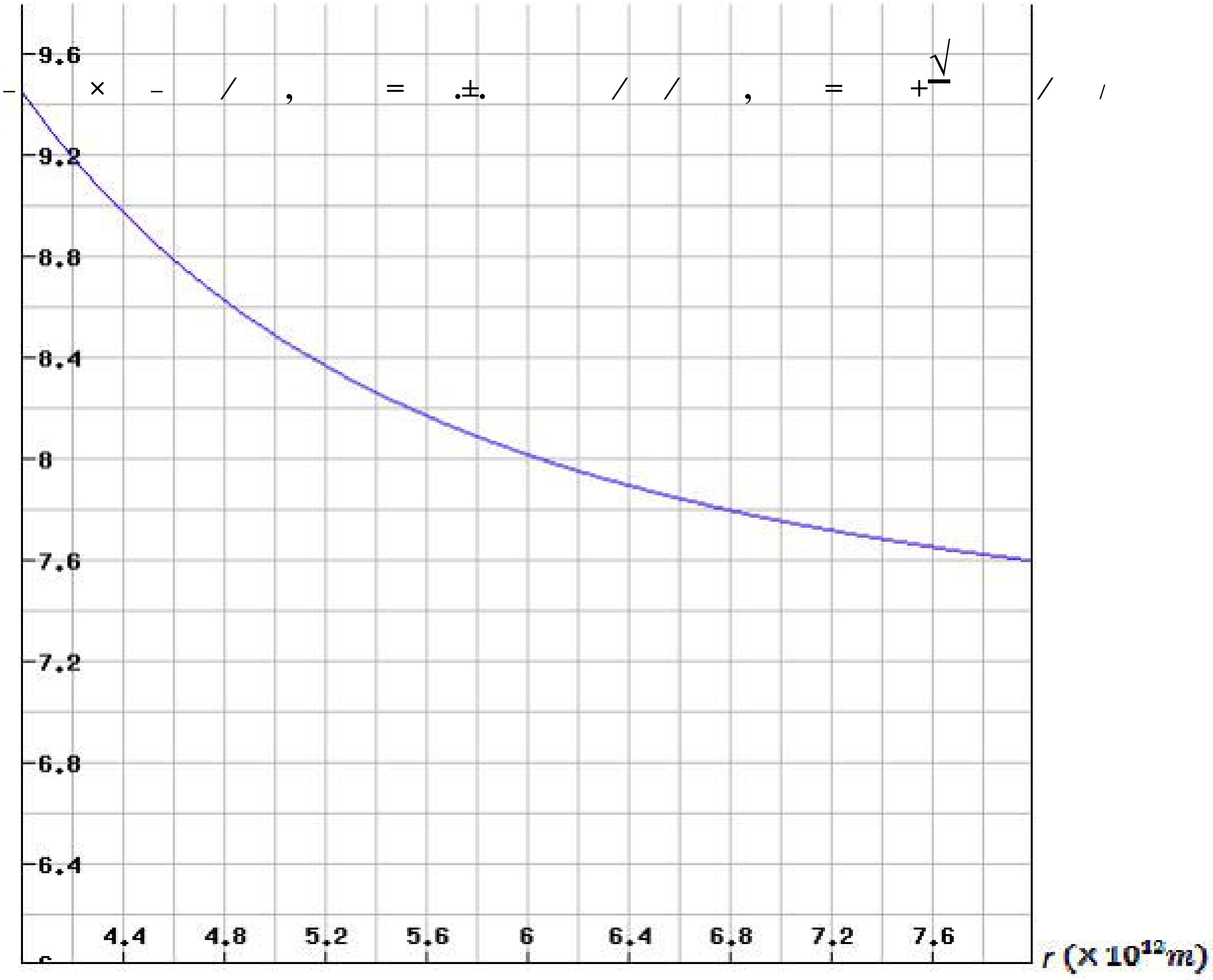


Fig. (3), the predicted Pioneer 10 anomaly versus distance from the Sun according to my solution.

References

[1] A. AlMosallami, “Reinterpretation of Lorentz transformation according to the Copenhagen school and the quantization of gravity,” Physics Essays, Volume 29: Pages 387-401, (2016).

[2] A. AlMosallami, Int. J. Modern Theor. Phys. 3, 44 (2014).

[3] A. AlMosallami, “A modified special relativity theory in the light of breaking the speed of light,” e-print viXra:1111.0001v1 [Relativity and Cosmology].

[4] A. AlMosallami, “The exact solution of the pioneer anomaly according to the general theory of relativity and the Hubble’s law,” e-print viXra:1109.0058 [Relativity and Cosmology].

[5] A. AlMosallami, IJSER 5, 128 (2014).

[6] A. AlMosallami, The Comalogical Theory (Toward a Science of

Consciousness, Arizona, Tucson 2008).

[7] Giuseppe Giuliani, “Experiment and theory: the case of the Doppler effect for photons,” e-print arXiv:1502.05736v1 [physics.hist-ph].

[8] W. Engelhardt, “Classical and Relativistic Derivation of the Sagnac Effect,” e-print arXiv:1404.4075v1 [physics.gen-ph].

[9] W. Engelhardt, “On the Origin of the Lorentz Transformation,” e-print arXiv:1303.5309v1 [physics.gen-ph].

[10] Matthias Sonnleitner, Nils Trautmann, and Stephen M. Barnett, Phys. Rev. Lett. 118, 053601, (2017).

[11] Anderson J. D. et al, “Study of the anomalous acceleration of Pioneer 10 and 11”, Physical Review D, V. 65, 082004.

[12] Anderson J. D. et al., “Indication, from Pioneer 10/11, Galileo, and Ulysses Data, of an Apparent Anomalous, Weak, Long-Range Acceleration,” arXiv:9808081v2 [gr-qc].

[13] S. G. Turyshev, arXiv:1107.2886v1

[14] C. B. Markwardt, “Independent Confirmation of the Pioneer 10 Anomalous Acceleration” arXiv:0208046v1 [gr-qc].

[15] V. T. Toth, Int. J. Mod. Phys. D 18, 717 (2009), arXiv:0901.3466.

[16] Riess, Adam G.; Lucas Macri, Stefano Casertano, Hubert Lampeitl, Henry C. Ferguson, Alexei V. Filippenko, Saurabh W. Jha, Weidong Li, Ryan Chornock (1 April 2011). "A 3% Solution: Determination of the Hubble Constant With the Hubble Space Telescope and Wide Field Camera". The Astrophysics Journal 730 (2).Bibcode 2011ApJ...730..119R. doi:10.1088/0004- 637X/730/2/119.

[17] Beutler, Florian; Chris Blake, Matthew Colless, D. Heath Jones, Lister Staveley-Smith, Lachlan Campbell, Quentin Parker, Will Saunders, Fred Watson (25 July 2011). "The 6dF Galaxy Survey: Baryon Acoustic Oscillations and the Local Hubble Constant". Monthly Notices of the Royal Astronomical S

[18] See wikipedia.org/wiki/Hermann_Minkowski#Work_on_relativity for “Hermann Minkowski,” Wikipedia.

[19] See wikipedia.org/wiki/Problem_of_time for “Problem of time,” Wikipedia.

Dark Matter Subhalo Interpretations Using Machine Learning: The Fourth Fermi-Lat Catalog

Amrutaa Vibho[1, 3], Rida Assaf [2]
[1] Guwahati, Assam, India, amrutaaghy2019@gmail.com (corresponding author)
[2] Department of Computer Science, University of Chicago, rida@uchicago.edu

Abstract

The quest for detecting dark-matter subhalos within the Galactic halo has taken many forms. Particularly interesting and promising is the use of spectral degeneracies to distinguish otherwise indistinguishable gamma-ray sources with near-null star formation. In further exploration of this realm, we attempt to classify high-latitude, non-variable, unassociated gamma-ray sources with Pulsar-like spectra in the 20-70 GeV Dark Matter annihilation range. Implementing supervised machine learning models on the 5788 gamma-ray sources recorded in the ten-year *Fermi*-LAT catalog (4FGL-DR2), where 1667 were formerly unassociated, we classify a total of 30 recorded gamma-ray events over a galactic latitude of 10 degrees, |b| >= 10 with a mean accuracy over 98%. This classification allows us to present a subset of potentially unanticipated gamma-ray sources as high-confidence Dark Matter Subhalo candidates.

Keywords: dark matter, astronomy, *Fermi*-LAT, subhalo, machine learning, gamma-rays, pulsars

1. Introduction:

Zwicky [1] postulated the existence of Dark Matter by using data that suggests there exists more matter in the universe than what is visible. Many other research endeavors point to the same, including but not limited to studies on rotation curves of spiral galaxies, large-scale structures, and hierarchical assembly processes.

Owing to recent scientific advancements in research methods, cold Dark Matter n-body simulations predict that the Milky Way halo should be heavily populated by the lightest Dark Matter substructures with limited or no star formation, which would be difficult to detect through optical surveys [2]. These Subhalos are therefore investigated best by tracking gamma-rays and annihilating Dark Matter. NASA's Large Area Telescope (LAT) onboard the *Fermi* satellite has provided scientists with the largest database for high-energy observations of stellar, galactic, and extragalactic structures and substructures through an incremental record of gamma-ray sources in the form of the FGL catalog [3, 4, 5]. Several approaches have been applied to classify the unassociated gamma-ray sources presented in the periodic FGL catalogs using Machine Learning. These include the classification of 4FGL gamma-ray sources into Pulsar and Active Galactic Nuclei [6], BL Lacs, and Flat-Spectrum Radio Quasars [7], and possibly Dark Matter Subhalo interpretations using formerly unassociated Pulsar candidates [8, 9].

Previous attempts to interpret these Subhalo structures have focused on classifying the gamma-ray events recorded in the FGL catalogs at higher (nearly extra)galactic latitudes – excluding any galactic sources such as supernova remnants and pulsar wind nebulae – that were formerly unassociated with a known gamma-ray source.

We approach the classification of the 4FGL unassociated sources with pulsar-like spectra or features using Artificial Neural Networks, with a secondary classifier of Support Vector Machines at a latitude 10 degrees above the galactic plane. In doing so, not only do we substantially reduce the search space for Dark Matter Subhalos but classify the candidates with the highest confidence and with multiple iterations of the model.

[3] Affiliation address - Amrutaa Vibho, 3B, Anil Nibash, Ganesh Mandir Road, Noonmati, Guwahati 781020, Assam, India

The paper is organized as follows: In Section 2 we describe the data used from the 4FGL catalog, and feature selection, along with the methodology adopted to prepare a training and classification set. We also introduce the software framework, followed by machine learning classifiers, Neural Networks, and Support Vector Machines. Section 3 presents results of training on and classifying the associated sources, then moving on to classification of the formerly unassociated gamma-ray sources from the catalog, with which we conclude after discussion in Section 4, acknowledgments in Sections 5 and 6, and references in Section 7.

2. Data and Methods:

2.1 Software Framework(s)

Our machine learning models were implemented and imported using scikit-learn, or *sklearn* [23].

For the purpose of this study, we made use of Google Colaboratory – simply known as Colab – a GPU-enabled open-source development tool under the TensorFlow AI framework owned by Google LLC [24], with the use of ML libraries Pandas and NumPy,using python as our coding language.

Given that the FGL catalogs are recorded as Flexible Image Transport System files (FITS), we imported the gamma-ray source catalog to the Colab environment making extensive use of astropy [25] – a collection of software packages enabling the use of python in handling astronomical data.

2.2 Dataset

For our study, we used the second data release, DR2 of the 4th FGL catalog, comprising an additional 2 years worth of gamma-ray events recorded to total 10 years' LAT observations [10], adding onto the initial data release DR1 which included 8 years' worth of gamma-ray events collected by LAT [11]. The Large Area Telescope onboard NASA's *Fermi* satellite was launched on June 11th, 2008, and has been surveying the sky every day. All gamma-ray events recorded are in the 50 MeV to 100 TeV energy range [3, 5].

The data release features more than 5000 gamma-ray sources where nearly one-fourth of the sources remain unassociated to known classes. We attempt to classify these unassociated sources into a binary classification of pulsars and extragalactic sources (Active Galactic Nuclei, Blazars, etc). The incremental 10-year catalog comprises 5788 sources, with 1667 unassociated.

In the .fits catalog file for 4FGL-DR2 – `gll_psc_v27` – the gamma-ray events recorded are divided into and categorized as three classes – associated (attributed to a known source), firmly identified (associated and their sources have been confirmed), and unassociated (sources are unknown and need association). Accordingly, the source types are capitalized to indicate firm identification (AGN, BCU, BLL, FSRQ, NLYS1, RDG, PSR), presented in lowercase for association (agn, bcu, bll, nlys1, rdg, ssrq, css, sey, glc, psr), and left blank as empty quotations to represent unassociated sources (''). Tables 1 and 2 list the number of associated, firmly identified, and unassociated sources as recorded in the catalog at the time of this research.

In table 1 that follows, we present the categorization using which the gamma-ray sources have been classified as *associated* classes, along with the number of sources known for each associated class in the catalog. We use these sources for our training sample in the machine learning models aimed at classifying unassociated sources.

Source Type	**Class**	**No. of sources**
Active Galactic Nuclei	AGN, agn	11
Blazar Candidates (uncertain)	BCU, bcu	1384
BL Lacs	BLL, bll	1308
Flat Spectrum Radio Quasars	FSRQ	744
Narrow-Line Seyfert 1	NLYS1, nlys1	9
Radio Galaxies	RDG, rdg	44
Soft Spectrum Radio Quasars	ssrq	2
Compact Steep Spectrum Quasar	css	5
Seyfert Galaxy	sey	1

Table 1 - Associated Sources in the 4FGL Catalog

Note: *FSRQs recorded in the catalog have been firmly identified and hence do not have a lowercase counterpart. On the contrary, ssrq, css, and sey have only been associated and not firmly identified and hence lack uppercase counterparts.*

Several studies and cosmological simulations strengthening the argument for the presence of dark matter also indicate that the Milky Way halo should be heavily populated with thousands of smaller dark matter subhalos, including but not limited to hosts for largest known dwarf spheroidal galaxies in the Milky Way and the lightest predicted dark matter substructures [8, 18, 19].

Since these substructures include any possible dark matter configuration, if they should exist, in concept, the bulk of this population is dominated by those substructures that nearly lack stellar components. In this scenario it would be impossible to detect dark matter subhalos through optical surveys alone. Spectral qualifications help resolve this problem.

Previous studies have brought to light two additional gamma-ray emitters (other than globular clusters hosting Millisecond Pulsars [8, 20]): dark matter subhalos [21] and dwarf galaxies [22]. Dark matter subhalos are expected to dominate gamma-ray emissions in this case.

Additionally, Weakly Interacting Matter Particles (WIMPs) and Massive Compact Halo Objects (MACHOs) remain the two most popular dark matter candidates. Early studies of unassociated gamma-ray sources recorded by the *Fermi*-LAT have recognized that WIMPs in the nearest subhalos could produce annihilation spectra nearly indistinguishable from that of gamma-ray pulsars. Thus, high-latitude, non-variable gamma-ray pulsar candidates without detected gamma-ray events also qualify as potential dark matter subhalo candidates. [8, 9, 12 - 17].

This brings us to include a subset of identified pulsars, globular clusters, and associated pulsars (as in table 2) as our subhalo training set. The candidates in our subhalo set fall within the 20-70 GeV range, making them consistent with the annihilation range of dark matter particles [9, 12].

In table 2, we gather the Pulsar Subhalo class along with the number of Pulsar gamma-ray sources per class.

Source Type	**Class abbreviation**	**No. of gamma-ray sources**
Identified Pulsars	PSR	232
Globular Clusters	glc	30
Associated Pulsar	psr	7

Table 2 - Unassociated sources with pulsar-like spectra or features present in 4FGL

2.3 Feature Selection

A large number of features introduces the need to determine their relative importance so that the basis for classification is as accurate as possible. Random Forests make this task easier since a higher value of Gini importance -- a Random Forest parameter -- corresponds to a greater role of the feature in classification [26, 27].

Unlike the many previous studies that target an issue beyond just the classification of unassociated gamma-ray sources [8, 9, 12], we propose to classify the unassociated sources using machine learning algorithms that have not been used or at least primarily, in the detection of dark matter. We attempt to experiment with this realm of research by further narrowing down the search space by using an almost extragalactic range of galactic latitude values. For these reasons, we build on the spectral prescriptions and feature selection of the latest paper that matches our aim in its first part [9], allowing us to focus primarily on the classification using Neural Networks and Support Vector Machine.

Accordingly, we use the following features as predictors from the *Fermi* catalog (selected in the original paper in line with the highest Gini parameter values obtained by them):

'LP_SigCurv' - Significance of fit improvement between PowerLaw and LogParabola

'PLEC_SigCurv' - Significance of fit improvement between PowerLaw and LogParabola for the PLSuperExpCutoff model

'Flux1000' - Photon Flux from 1 Gev to 100 Gev (integral)

'LP_beta' - Curvature parameter for fitting LogParabola

'Frac_Variability' - Fractional Variability in yearly detected fluxes

'Variability_Index' - Index (difference) of flux fitted per time interval and average flux over the full-time interval of the catalog

'PLEC_Index' - Low-energy photon index for fitting PLSuperExpCutoff

'PL_Index' - Photon index for PowerLaw fit

'Pivot_Energy' - Minimal differential flux (error) energy

'LP_Index' - Photon index at Pivot Energy for PowerLaw fit

2.4 Machine Learning Classifiers

Inspired by the many previous ML approaches on the FGL catalog and given the one of particular interest to us [9] wherein they implemented Random Forests and XGBoost, we too began with that implementation to understand the arrangement and organization of the 4FGL-DR2 and the source-association structure. Like the aforementioned study, we too found Random Forests to perform at an accuracy of 97.7% and XGBoost 97.4% for |b| >= 10, which is no different from their result.

We decided on approaching our binary classification problem with Neural Networks and provide a comparative measure with Support Vector Machine. This choice is motivated primarily by the fact that this combination of algorithms hasn't previously been implemented on the FGL catalogs for dark matter subhalo classification, though it has been used in the catalog for classifying other active galactic nuclei and gamma-ray sources [7].

Neural Networks

MLPClassifier is a perceptron-based algorithm that implements classification by *learning* the underlying connection between perceptrons, thereby forming an artificial neural network. The most important part of training an MLP model is that as a multi-layer-perceptron, the number of neural network layers between the input and the output is greater than one, also hidden.

Hidden layers are the most important parameter for MLPClassifier and hyperparameter tuning just about decides the end result, which is why we treated this part with caution, in particular, the anticipated risks of overtraining.

Previous studies not just limited to the FGL catalogs but those that deal with Machine Learning models and implementations have suggested the use of hyperparameter optimizers RandomSearchCV and GridSearchCV. Given the complexity of our model, we decided to use the systematic and not random hyperparameter tuning method using GridSearchCV. Besides this, as mentioned above, owing to the complexity of our model we were aware that overtraining may produce a skewed result subset and therefore decided on implementing K-Fold Cross-Validation.

Usually, models are either optimized for hyperparameters or cross-validated; the order of implementation is predetermined. But some models, especially the more complex ones like ours, are worked on through both, simultaneously. Now while a plethora of techniques exist for getting the best ML model possible, the time and cost imposed by the execution put significant constraints. Therefore, to get the best of both optimizations and also maintain ML efficiency, we implemented Nested K-Fold Cross-Validation.

In Nested or Double Cross-Validation, model hyperparameter optimization is treated as a part of the model itself, evaluating it within the broader k-fold cross-validation procedure for model comparison and selection.

Our confidence in these approaches was confirmed, when we achieved a high accuracy with a not so deep layer size for MLP, thus significantly preventing overtraining, as further discussed in Section 3.

Support Vector Classifier

The reason we chose SVC for the task of cross-referencing, is because SVC is non-linear and handles complex classifications better than standard decision trees and linear models. Even so, we treated this model with equal caution and care as MLP in respect of the risk of overtraining and hyperparameter optimization. For SVC we used Repeated Stratified K-Fold Cross-Validation along with BayesSearchCV, a part of scikit-optimize or *skopt.*

Bayesian Optimization provides a principled technique based on Bayes Theorem to direct a search of a global optimization problem. It is believed to offer an efficient alternative to the comparatively less efficient hyperparameter optimization procedures such as GridSearchCV and RandomSearchCV.

Our confidence in these approaches was confirmed, when we achieved a high accuracy and an optimum value for the regularization parameter for SVC , thus significantly preventing overtraining, as further discussed in Section 3.

2.5 Training and Testing

Galactic Latitude

To avoid contamination from galactic sources, we only use gamma-ray events at latitude of ten degrees away from the Milky Way's plane, i.e., |b| >= 10. Note that we use absolute values to avoid directional negatives above and below the plane.

Missing values and building the training dataset

After applying the constraints for classification features and galactic latitudes, we detected missing values in the dataset. To move forward, we eliminated all null values, resulting in a smaller dataset of 4118 sources above |b| = 10.

Class Imbalance

Upon gathering a dataset with complete data, we realized the need to counter class imbalance, since in both the data releases, Pulsars form less than ten percent of the training and testing dataset, largely dominated by the extragalactic sources. Machine learning classifiers tend to be biased towards the majority class, and this could result in misclassification of crucial sources.

First, we set the split ratios for training and testing to 70-30 respectively. Then, beginning with RandomOverSampler, we implemented oversampling techniques that could artificially generate Pulsar candidates for the purpose of training the dataset in a way that both labels contain near-equal instances.

ADASYN [37] uses density distribution in automatically deciding the number of synthetic samples per minority sample adaptively changing weights of different minority samples. Whereas, SMOTE [36] generates the same number of synthetic samples for each original minority sample to compensate for skewed distribution. Additionally, SMOTE uses K-nearest neighbors, i.e., it selects k nearest neighbors and joins them to further create synthetic samples.

Given the fundamental workability of SMOTE and relative simplicity, given that we are dealing with a complex feature set and KNN's efficiency with classification problems, out of RandomOverSampler, Adaptive Synthetic Sampling (ADASYN), and Synthetic Minority Oversampling (SMOTE), SMOTE better suited our purpose of attributing similar artificial sources for the Pulsar class.

We also initially implemented RandomOverSampling on the minority class, increasing it by 10% percent of the majority class size, and then using RandomUnderSampling on the majority class, to downsize it and make both labels nearly equal. This seemed to eliminate a lot of the necessary sources from the majority class, hence we decided to proceed with SMOTE on the minority, increasing it by 35% of the majority class size.

3. Results:

Training and testing on the associated classes

Upon eliminating potential factors contributing towards misclassification (through SMOTE), we trained our MLP and SVC models on the sources already associated with known classes, as presented in the 4FGL catalog(s).

For MLP, we used the cross_val_score() function that executed the outer cross-validation procedure on the configured GridSearchCV, automatically using the refit best performing model on the test set from the outer loop. With 5 inner and 10 outer splits (or folds), we obtained a mean Accuracy: 0.981 (0.006), while the best parameter determined was a hidden layer size of (10,).

For SVC, we implemented hyperparameter tuning and RepeatedStratifiedKFold Cross-Validation using 10 splits (or folds), and 3 repetitions, obtaining a best score of 0.983 and the best value for the regularization parameter or C as 100.

Classifying the unassociated sources

After training the models on sources associated with known classes, we began implementing the same strategy to deal with missing values and applying galactic latitude constraints on the unassociated classes. The *clean* dataset of unassociated sources that we classified, finally consisted of 835 gamma-ray sources at latitude greater than ten degrees.

Since our dataset was trained on models that qualify unassociated sources with pulsar-like spectra or features and that match the spectrum of Dark Matter annihilation, with our narrow search space, we report a subset of 30 high-confidence potential dark matter subhalo candidates among the 5788 gamma-ray sources recorded in the *Fermi*-LAT catalog with a probability of classification greater than 70%. To understand and motivate our selection of the results, we present plots correlating True and False Positives for MLP and SVC through the Receiver Operating Characteristic (ROC) curve in figures 1 and 2 respectively.

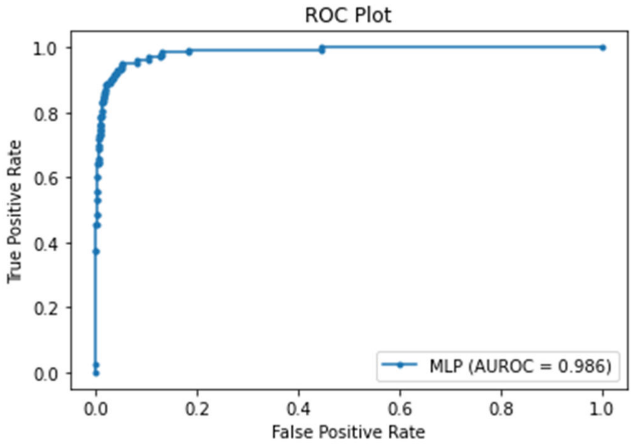


Figure 3 presents a comparative plot of ROC for MLP and SVC, followed by table 3, introducing Dark Matter Subhalo candidates classified in the 4FGL-DR2 catalog at |b| >= 10.

Figure 1: The ROC curve correlating True and False Positive Rates for MLP, with an AUROC score of 0.986

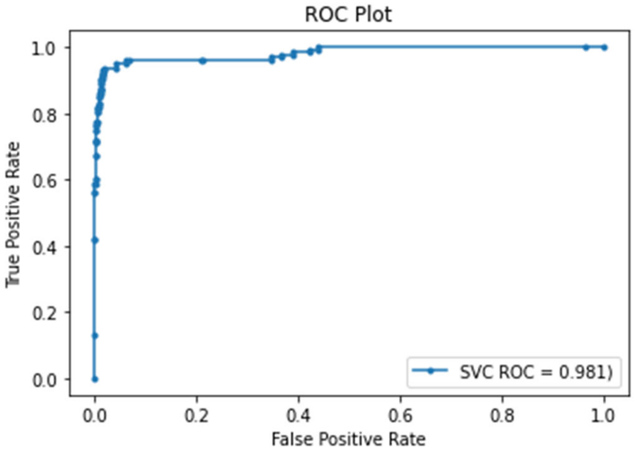


Figure 2: Figure 1: The ROC curve correlating True and False Positive Rates for SVC, with an AUROC score of 0.981.

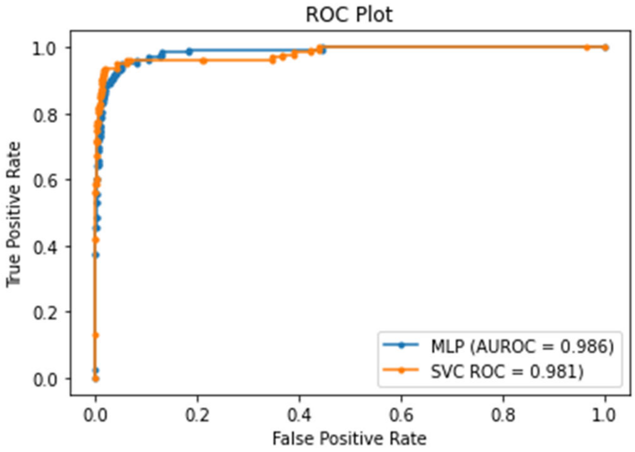


Figure 3: Plot with ROC curves for MLP and SVC, correlating their respective True and False Positive Rates

Source_Name	**Right Ascension**	**Declination**
4FGL J0026.6-4600	6.6521	-46.0156
4FGL J0034.3-0534	8.5961	-5.5809
4FGL J0035.0-5728	8.7521	-57.4714
4FGL J0043.6+2223	10.9095	22.3864
4FGL J0047.1-6203	11.777	-62.052
4FGL J0102.9-7051	15.739	-70.863
4FGL J0111.4+0534	17.8573	5.5761

4FGL J0114.9-3400	18.7398	-34.0063
4FGL J0116.2-6153	19.06	-61.8947
4FGL J0134.3-3842	23.5887	-38.7085
4FGL J0142.7-0543	25.6754	-5.7332
4FGL J0143.5-3156	25.8977	-31.9467
4FGL J0150.9+1230	27.7272	12.5135
4FGL J0151.7+5455	27.9312	54.9172
4FGL J0204.3-3140	31.0751	-31.6713
4FGL J0212.9+2244	33.2427	22.7466
4FGL J0221.2-1312	35.313	-13.2046
4FGL J0221.8+3730	35.469	37.511
4FGL J0221.5+2513	35.3809	25.2305
4FGL J0224.2+1616	36.0681	16.2667
4FGL J0240.2-0248	40.0534	-2.8086
4FGL J0301.4-3124	45.351	-31.4103
4FGL J0242.6-0000	40.667	-0.0069
4FGL J0325.9-1843	51.4839	-18.725
4FGL J0330.7-2408	52.6938	-24.147
4FGL J0334.2-4008	53.5566	-40.145
4FGL J0339.5-0146	54.8771	-1.7769
4FGL J0347.0-6400	56.7604	-64.0034
4FGL J0350.0+0640	57.5043	6.6754
4FGL J0352.0-2516	58.0034	-25.2733

Table 3 - High confidence classifications for potential Dark Matter Subhalo candidates in 4FGL-DR2 (30), with their right ascension, and declination.

Figures 4 and 5 show the locations of the candidates in RA-DEC Galactic sky coordinates for MLP and SVC respectively.

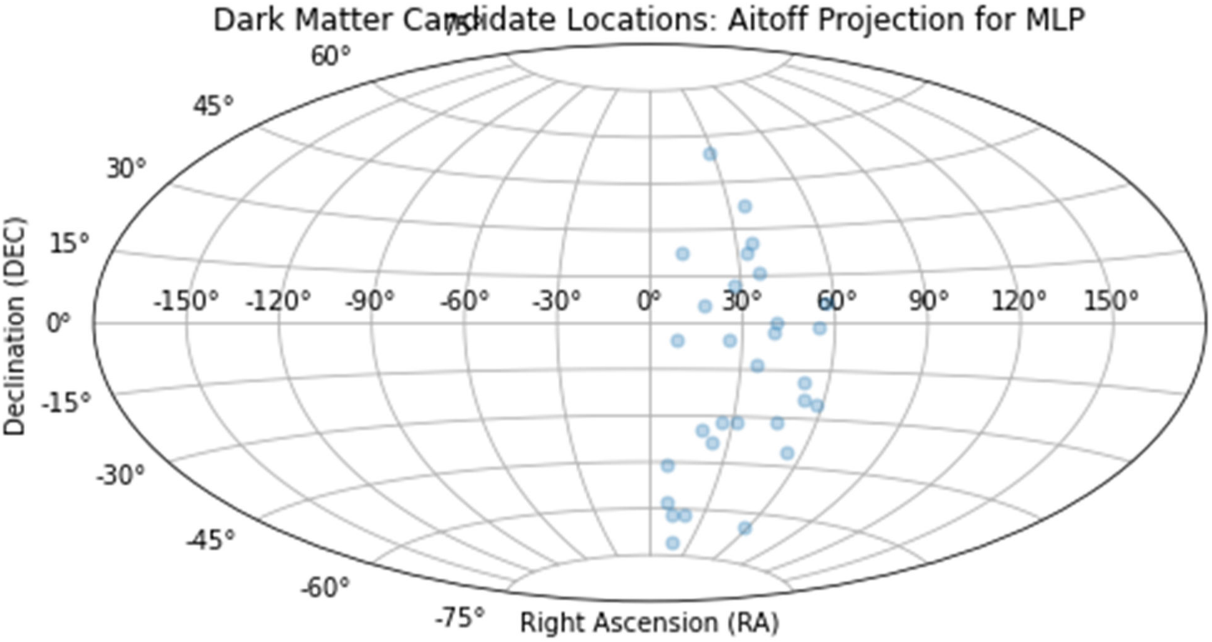


Figure 4: Locations of 30 Dark Subhalo candidates as an Aitoff Projection of RA/DEC (blue circles) for MLP

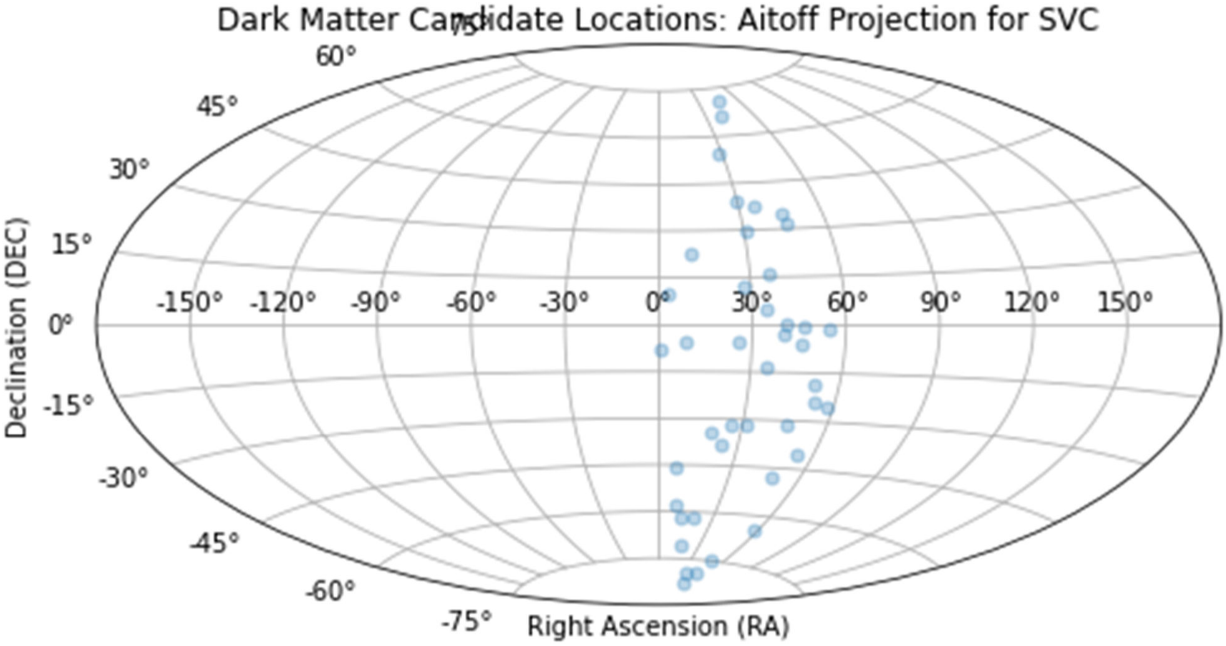


Figure 5: Locations of 45 Dark Subhalo candidates as an Aitoff Projection of RA/DEC (blue circles) for SVC

4. Discussion:

Previous studies that motivated our paper

Machine learning has applications in nearly every field of research today. Our study is motivated by the staggering pace at which astrophysical and astronomical research is yielding results and the fact that ML could further this pace with the scientific community we have. During the process of finalizing our objective and issue of study, we came across innumerous papers using ML in astrophysical research [32-25], and it is only after going through their remarkable strategies and methods that we landed our sights on DM Subhalo classification using ML. It is noteworthy to mention that apart from DM Subhalo classification, ML has been used for various other purposes on the *Fermi*-LAT catalogs and we think our paper would be incomplete without mentioning a few of them that inspired us for our study [28-31].

What is new through our paper

In a previous study done on the 4FGL, 73 formerly unassociated gamma-ray sources were reported as potential Dark Matter Subhalo candidates at a galactic latitude above 10 degrees [9]. Before that, in 3FGL, the same was true for 34 high-confidence sources at a galactic latitude above 5 degrees [8].

Through this paper, we approached the classification of the 4FGL unassociated sources with pulsar-like spectra or features using Artificial Neural Networks, with a secondary classifier of Support Vector Machines at an almost extragalactic latitude range, ten degrees above the galactic plane. These Machine Learning algorithms are implemented in this paper as a few of the firsts in using ANNs in Dark Matter detection using the FGL catalog. With cross-validated and tuned models, we present a concentrated and high-confidence subset of 30 Dark Matter Subhalo candidates.

Our results differ from the previous ones in that between the time the previous studies were published and ours was undertaken, some formerly unassociated gamma-ray sources were classified as pulsars and moved into the unassociated classes.

Conclusions and future work

Unassociated sources with pulsar-like spectra or features are related to a non-luminous spectrum. The fact that our training label sources matched spectra of dark-matter annihilation further pinpoints the relevance of these high-confidence candidates when classifying unassociated sources using the trained model.

We believe that the 30 high-confidence potential Dark Matter Subhalo candidates common to MLP and SVC could add value to the existing set of Dark Matter candidates, and with future increments in the FGL catalog, the total number of candidates may increase substantially. Meanwhile, we do acknowledge that the additional candidates classified by SVC remain objects for further investigation, since numbers higher than 30 have already been accomplished in other papers using different algorithms. The additional candidates are given as follows:
4FGL J0001.2-0747, 4FGL J0013.4+0950, 4FGL J0202.4+2943, 4FGL J0216.8+0510, 4FGL J0224.0-7941, 4FGL J0239.1+6634, 4FGL J0250.2-8224, 4FGL J0257.8+7044, 4FGL J0300.4+3450, 4FGL J0303.2+3149, 4FGL J0303.3-7913, 4FGL J0304.5-0054, 4FGL J0304.9-0606, 4FGL J0313.6-7508, 4FGL J0314.4-4805.

In part, our belief is sustained by the knowledge that a narrow search space leaves more room for investigation than contamination of sources. This further indicates that with future increments in the FGL, considering higher galactic latitudes could provide a larger investigation base into the spectral and physical characteristics once Dark Matter candidates are classified particularly for a closer look into stellar stream perturbations [9], millisecond pulsars [8], and other interpretations of these sources.

5. CRediT Author Statement (Contributor Roles Taxonomy):

Amrutaa Vibho: Conceptualization, Methodology, Data curation, Software, Writing - Original draft, editing, and final draft, Formal Analysis
Rida Assaf: Validation, Supervision, Project Administration, Writing - Review

6. Acknowledgments:

This study was facilitated by the mentorship of Mr. Rida Assaf, a Ph.D. candidate, and Machine Learning specialist at the University of Chicago, along with academic support through the Horizon Academic Research Program, for which I am immensely grateful.

During the course of this three-month-long endeavor, my knowledge of the Universe was enriched at Yale Summer Session in Stellar and Galactic Astronomy, Stanford University (Contd.) in high-energy astrophysics, and at the time of writing this paper, my enthusiasm for the skies has been enriched by an acceptance at the International Institute for Astronautical Sciences through Project PoSSUM for Polar Suborbital research in the Upper Mesosphere.

It is noteworthy that this paper is a result of collaborative inspiration from the works of citations listed below, and could not have been possible without the resources made available through NASA GSFC-Fermi, sklearn, and Google Colaboratory.

Lastly, I express my most sincere gratitude to my parents and twin sister for their continued support despite the many tribulations through the course of this project. I owe my perseverance to them.

7. References:

[1] Zwicky, 1933. The Redshift of Extragalactic Nebulae. *Helvetica Physics Acta, Vol 6, p. 110-127.* https://arxiv.org/ftp/arxiv/papers/1711/1711.01693.pdf

[2] Hezaveh et al., 2012. Dark Matter Substructure Detection using Spatially Resolved Spectroscopy of Lensed Dusty Galaxies *The Astrophysical Journal Supplement Series, 247:33 (37pp).* https://iopscience.iop.org/article/10.1088/0004-637X/767/1/9/pdf

[3] Abdollahi et al., 2020, Fermi Large Area Telescope Fourth Source Catalog. *The Astrophysical Journal Supplement Series, 247:33 (37pp).* Fermi Large Area Telescope Fourth Source Catalog

[4] The Fermi-LAT collaboration (2020), Fermi Large Area Telescope Fourth Source Catalog. *Astrophysical Journal Supplement, 247, 33.* http://hdl.handle.net/21.11116/0000-0008-1C35-E

[5] Ballet et al., 2020, *Fermi* Large Area Telescope Fourth Source Catalog Data Release 2, https://arxiv.org/pdf/2005.11208.pdf,

[6] Zhu K. R. et al., 2020, Searching for AGN and Pulsar Candidates in 4FGL Unassociated Sources Using Machine Learning. *Research in Astronomy and Astrophysics, July 1:15*
https://arxiv.org/pdf/2001.06010.pdf

[7] Germani S. et al., 2021, Artificial Neural Network Classification of 4FGL Sources, *Monthly Notices of the Royal Astronomical Society, Volume 505, Issue 4, August 2021, Pages 5853–5861.* https://doi.org/10.1093/mnras/stab1748

[8] Mirabal N. et al, 2016, 3FGL DEMOGRAPHICS OUTSIDE THE GALACTIC PLANE USING SUPERVISED MACHINE LEARNING: PULSAR AND DARK MATTER SUBHALO INTERPRETATIONS, *The Astrophysical Journal, 825:69 (8pp)* https://iopscience.iop.org/article/10.3847/0004-637X/825/1/69/pdf

[9] Mirabal, Bonaca, 2021, Machine-Learned Dark Matter Subhalo Candidates in the 4FGL-DR2: Search for the Perturber of the GD-1 Stream. https://arxiv.org/pdf/2105.12131.pdf

[10] LAT 10-year Source Catalog (4FGL-DR2)

[11] LAT 8-year Source Catalog (4FGL)

[12] Bertoni et al., 2015, Examining The Fermi-LAT Third Source Catalog In Search Of Dark Matter Subhalos, *Journal of Cosmology and Astroparticle Physics JCAP12(2015)035*, https://arxiv.org/pdf/1504.02087.pdf

[13] Edward A. Baltz et al., 2007, CAN ASTROPHYSICAL GAMMA-RAY SOURCES MIMIC DARK MATTER ANNIHILATION IN GALACTIC SATELLITES?, *The Astrophysical Journal, ApJ* 659 L125, https://iopscience.iop.org/article/10.1086/517882/pdf

[14] Buckley and Hooper, 2010, Dark Matter Subhalos in the First Fermi Source Catalog, *Physical Review D, vol. 82, Issue 6, https://arxiv.org/pdf/1004.1644.pdf*

[15] Nieto et al., 2011, A search for possible dark matter subhalos as IACT targets in the First Fermi-LAT Source Catalog, *https://arxiv.org/pdf/1110.4744.pdf*

[16] Ackermann et al., 2012, Search for Dark Matter Satellites using *Fermi*-LAT, *The Astrophysical Journal, Volume 747, Number 2, M. Ackermann et al 2012 ApJ 747 121, https://iopscience.iop.org/article/10.1088/0004-637X/747/2/121*

[17] Belikov et al., 2012, Searching for Dark Matter Subhalos in the *Fermi*-LAT Second Source Catalog, *Physical Review D, Volume 96, Issue 4, Phys. Rev. D 86, 043504 (2012) - Searching for dark matter subhalos in the Fermi-LAT second source catalog*

[18] Ricotti et al., 2009, A New Probe of Dark Matter and High-Energy Universe using Microlensing, *The Astrophysical Journal, Volume 707, Number 2,* https://iopscience.iop.org/article/10.1088/0004-637X/707/2/979/pdf

[19] Scott et al., 2009, Gamma Rays from Ultracompact Primordial Dark Matter Minihalos, *Physical Review Letters, Volume 103, Issue 21,* https://journals.aps.org/prl/pdf/10.1103/PhysRevLett.105.119902

[20] Acero et al., 2015, *FERMI* LARGE AREA TELESCOPE THIRD SOURCE CATALOG, *The Astrophysical Journal Supplement Series, Volume 218, Number 2,* https://iopscience.iop.org/article/10.1088/0067-0049/218/2/23/pdf

[21] Bergstrom et al., 1999, Clumpy Neutralino Dark Matter, *Physical Review D, Volume 59, Issue 4, Phys. Rev. D 59, 043506 (1999) - Clumpy neutralino dark matter*

[22] Lake et al., 1990, Detectability of gamma-rays from Clumps of Dark Matter, *Nature 346, 39–40 (1990), Detectability of γ-rays from clumps of dark matter | Nature*

[23] Pedregosa et al., 2011, Scikit-learn: Machine Learning in Python, *Journal of Machine Learning in Research, JMLR 12, pp. 2825-2830, 2011.*

[24] Google Colaboratory, https://colab.research.google.com/notebooks/gpu.ipynb
[25] Astropy, https://docs.astropy.org/en/stable/index.html

[26] Ceriani, L., Verme, P, 1912, The origins of the Gini index: extracts from Variabilità e Mutabilità, *The Journal of Economic Inequality 10, 421–443 (2012),* https://doi.org/10.1007/s10888-011-9188-x

[27] Liaw and Wiener, 2002, Classification and regression by randomforest, *R News 2, 18–22,* https://www.r-project.org/doc/Rnews/Rnews_2002-3.pdf

[28] Saz Parkinson et al., 2016, CLASSIFICATION AND RANKING OF FERMI LAT GAMMA-RAY SOURCES FROM THE 3FGL CATALOG USING MACHINE LEARNING TECHNIQUES, *The Astrophysical Journal,* ApJ 820: 8, https://iopscience.iop.org/article/10.3847/0004-637X/820/1/8/pdf

[29] Shi-Ju Kang et al., 2019, Evaluating the Classification of Fermi BCUs from the 4FGL Catalog Using Machine Learning, *The Astrophysical Journal, ApJ* 887 134, https://iopscience.iop.org/article/10.3847/1538-4357/ab558b/pdf

[30] Bhat and Malyshev, 2021, Machine learning methods for constructing probabilistic Fermi-LAT catalogs, https://arxiv.org/pdf/2102.07642.pdf

[31] Einecke S, 2016, Search for High-Confidence Blazar Candidates and Their MWL Counterparts in the Fermi-LAT Catalog Using Machine Learning, *Galaxies 4(3), 14,* https://www.mdpi.com/2075-4434/4/3/14/htm

[32] Graf et al., 2014, SkyNet: an efficient and robust neural network training tool for machine learning in astronomy, *Monthly Notices of the Royal Astronomical Society*, Volume 441, Issue 2, 21 June 2014, Pages 1741–1759, https://academic.oup.com/mnras/article/441/2/1741/1071156?login=true#19331055

[33] Gao and Zhang, 2020, Deflating Super-puffs: Impact of Photochemical Hazes on the Observed Mass–Radius Relationship of Low-mass Planets, *The Astrophysical Journal, ApJ* 890 93,
https://iopscience.iop.org/article/10.3847/1538-4357/ab6a9b/meta

[34] Sotin et al., 2007, Mass-radius curve for extrasolar Earth-like planets and ocean plants, *Icarus (Elsevier), Volume 191, Issue Issue 1, Pages 337-351,* https://www.sciencedirect.com/science/article/abs/pii/S0019103507001601

[35] Aurisano et al., 2016, A convolutional neural network neutrino event classifier, *Journal of Instrumentation, Volume 11, JINST 11 P09001*, https://iopscience.iop.org/article/10.1088/1748-0221/11/09/P09001/pdf

[36] Chawla et al., 2011, SMOTE: Synthetic Minority Over-sampling Technique, *arXiv e-prints* https://arxiv.org/pdf/1106.1813.pdf

[37] He et al., 2008, ADASYN: Adaptive synthetic sampling approach for imbalanced learning, *IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 2008, pp. 1322-1328, https://ieeexplore.ieee.org/abstract/document/4633969

Emily Dickinson's Green Aesthetics: Mode Gakuen Cocoon Tower as the Anthropomorphic Architectural Representation in the Age of Anthropocene

Chia-Wen Kuo

***Abstract*—** Jesse Curran states that there is a "breath awareness" that "facilitates a present-minded capability" to catalyse an "epistemological rupture" in Emily Dickinson's poetry, particularly in the age of Anthropocene. In Dickinson's "Nature", non-humans are subjectified as nature ceases to be subordinated to human interests, and Dickinson's Eco-humility has driven us, readers, into mimicking nature for the making of a better world. In terms of sustainable architecture, Norman Foster is among the representatives who utilise BIM to reduce architectural waste while satiating the users' aesthetic craving for a spectacular skyline. Notably, the Gherkin - 30 St. Mary Axe in east-end London. In 2019, Foster and his team aspired to savour the London skyline with his new design - the Tulip, which has been certified by the LEED as a legitimate green building as well as a complementary extension of the Gherkin. However, Foster's proposition had been denied for numerous times by the mayor Sadiq Khan and the city council as the Tulip cannot blend in the public space around while its observatory functions like a surveillance platform. The Tulip, except for its aesthetic idiosyncrasy, fails to serve for the public good other than another ostentatious tourist attraction in London. The architectural team for Mode Gakuen Cocoon tower, completed in 2008, intended to honour Nature with the symbolism in the building's aesthetic design. It serves as an architectural cocoon that nurtures the students of "Special Technology and Design College" inside. The building itself turns into a Dickinsonian anthropomorphism, where humans are made humble to learn from the entomological beings for self-betterment in the age of Anthropocene. Despite bearing resemblance to a tulip as well as its LEED credential, Norman Foster's Tulip merely pays tribute to the Nature in a relatively superficial manner without constructing an apparatus that substantially benefit the Londoners as all green cities should embrace Emily Dickinson's "breath awareness" and be built and treated as an extensive as well as expansive form of biomimicry.

***Keywords*—** green city, sustianable architecture, London, Tokyo.

Chia-Wen Kuo is with the National Chengchi University, Taiwan (e-mail: fluorescentgazes@gmail.com).

Drivers and Barriers of Asphalt Rubber in Sweden

Raheb Mirzanamadi, João Patrício

Abstract**—**Asphalt rubber (AR) was initially developed in Sweden in the 1960s by replacing crumb rubber (CR) as aggregates in asphalt pavement. The AR produced by this method had better mechanical properties than conventional asphalt pavement but was very expensive. Since then, different technologies and methods are developed to use CR in asphalt pavements, including blending CR with bitumen with a high temperature in the mixture, called wet method, and blending CR with bitumen in the refinery, called terminal blending method. In 2006, the wet method was imported from the USA to Sweden to evaluate the potential of using AR in Swedish roads. 154 km AR roads were constructed by the wet method in Sweden. The evaluation showed that the AR had, in most cases, better mechanical performance than conventional asphalt pavements. However, the terrible smoke and smell led the Swedish Transport Administration (STA) stopped using AR in Sweden. Today, there are few focuses on AR, despite its good mechanical properties and environmental aspects. Hence, there is a need to study the drives and barriers of using AR mixture in Sweden. The aims of this paper are: (i) to study drivers and barriers of using AR pavements in Sweden and (ii) to discover knowledge gaps for further research on this area. The study was done using literature review and completed by interviews with experts, including three researchers from Swedish National Road and Transport Research Institute (VTI) and two experts from STA. The results showed that AR can be an alternative not only for conventional asphalt pavement but also for Polymer Modified Asphalt (PMA) due to the same mechanical properties but the lower cost for production. New technologies such as terminal blending and using Warm Mix Asphalt (WMA) methods can lead to reducing the energy and temperature during production processes. From this study, it is found that there is not enough experience and knowledge about AR in Sweden, and more research is needed, including: lifespan of AR, mechanical properties of AR using new technologies, and impact of AR on spreading and leaching substances into nature. More studies can lead to standardization of using AR in Sweden, a potential solution for the use of end-of-life tyres, with better mechanical properties and lower costs, in comparison with conventional asphalt pavements and PMA.

Keywords**—** Asphalt rubber, crumb rubber, terminal blending method, wet method

I. Introduction

End-of-life (EOL) tyres are characterized for having unique characteristics. Elastic modulus, water absorption, and tensile strength are just some examples, that make them a great resource to be employed in multiple applications [1]–[3]. Waste tyres are a large waste flow in Sweden. Considering the volume of generated EOL tyres from 2016 to 2019, an increasing trend from 83.774 tons to 94,550 tons is reported [5]. In 2020, 84,574 ton of tyres were generated in Sweden [4]. Currently, in Sweden, 65% of EOL tyres are used as an energy source for energy production, or in the cement industry, and the remaining 35 % are recycled as blasting mats [1].

The European Union highly recommends the development of strategies and processes that promote a circular economy [1], [6] If circular economy practices are well implemented, they can lead to several environmental gains, including the reduction of resources extracted from nature, and the reduction of the energy used in the production phase [7] For example, EOL tyres can contribute significantly to the circular economy if innovative recycling uses are applied. One possibility is to use Crumb rubber (CR) from EOL tyres to produce asphalt rubber (AR) in order to beneficially improve the properties and performance of asphalt mixtures [8].

AR can improve the resistance of asphalt pavements to rutting and fatigue damages and thereby can reduce maintenance and operation costs of the pavements [9]. There are three main processes of using CR to produce AR, namely: wet process (ARwet), dry process (ARdry), and terminal blending process (ARtb) [8].

Table 1 shows a summary of the different processes and technologies for recycling CR to produce AR mixtures. It should be noted that some researchers classified the terminal blending process as wet process [10], [11]. However, in this study, ARwet and ARtb are investigated separately.

Rubber modified asphalt concrete (RUMAC, also called "Skega Asphalt", "Rubit" or "Rubtop" in Sweden) was initially started in Sweden in the late 1960s [11]. RUMAC was the first ARdry in Sweden and the idea behind it was to replace a small portion of aggregates in asphalt pavements with the same fraction of CR. The content of CR was between 1% and 3% of the total aggregate weight and the air void content was between 2% and 4% [16]. The main goal of developing RUMAC asphalt was to improve road safety during winter seasons in Sweden, due to its higher resistance against ice formation. RUMAC asphalt was often used as wearing layers on bridge pavement structures to prevent ice formation. Furthermore, RUMAC provides specific wear and ruttning characteristics than conventional asphalt pavement [17]. However, the production and construction cost of RUMAC asphalt was almost twice that of conventional asphalt pavements, which led to stop using it in a wide range in Sweden [18].

In 2006, SAT initiated Swedish asphalt rubber development project to evaluate the potential implementation of ARwet concepts on Swedish roads. The aims of the project were [19]:

- To increase the lifespan of pavement and thereby to reduce annual life cycle costs.
- To reduce particle emissions and noise, occurred due to tyre and road interaction.
- To increase the friction between tyre and wet roads.

Raheb Mirzanamadi is with the Swedish National Road and Transport Research Institute, Sweden (e-mail: raheb.mirzanamadi@vti.se).

TABLE 1.
A SUMMARY OF THE DIFFERENT PROCESSES AND TECHNOLOGIES FOR RECYCLING CRUMB RUBBER TO PRODUCE ASPHALT RUBBER MIXTURES [8]–[10], [12]–[15]

	Wet process	Dry process	Terminal blending process
Process	Crumb rubber (CR) is blended with bitumen in an on-site mixing tank at a temperature between 176 °C and 226 °C for 45 min to 4 h, depending on the interaction process. The blend consists of a minimum 15% CR by weight of total binder and may include additives to help workability There are different methods for wet processes such as McDonalds, continuously blending-reaction system, filed blend and semi-wet process. Properties of ARwet can be influenced by blending method, agitation energy, blending temperature and time and CR size and amount	CR is blended with heated aggregates at a batch plant mixer at ambient temperature for a certain time and then bitumen is added to the mixer to produce ARdry. The interaction between CR and bitumen should be at least 90 min. CR is considered as a part of aggregate and its content can vary between 3% to 20% by weight of the total mixture. There are different methods for dry processes such as RUMAC, generic dry process, and chuck asphalt rubber. ARdry properties can be influenced by CR size, binder and air content and aggregate gradation.	CR is blended with a hot binder at the refinery or a stationary asphalt terminal. The mixture is heated for 2 h to 8 h to fully digest CR in the asphalt binder. Traditionally, ARtb consists of a maximum 10% CR. Today, the CR content can reach to 25%, if CR size is smaller than 0.6 mm. ARtb has high stability and storage life, and its manufacture is similar to PMA ARtb properties are more influenced by production temperature than interaction time. However, a temperature greater than 260 °C should be avoided.
Material	The maximum CR particle size is 2 mm and the average particle size is 0.56 mm. The most common asphalt type is 50/70 penetration depth. ARwet is mainly used for gap-graded and open-graded, less used for dense-graded binders due to low air void content WAR-ARwet and WAR-RAP-ARwet can decrease emission and energy consumption and cost.	Maximum CR particle size is 1 mm or even recommended to be smaller than 0.2 mm. The increase in the CR content can affect the air voids configuration and reduce the mechanical performance of ARdry. Gap-graded or dense-graded aggregates are preferred Higher binder content should be used, compared to HMA (1% to 2%) Binder with the same or higher penetration grade should be used, compared to HMA	The maximum CR particle size is 0.6 mm. Dense-graded aggregates are preferred. WAR-ARtb and WAR-RAP-ARtb and SMA-ARtb can decrease emission and energy consumption and cost.
Performance	ARwet has more flexibility due to swelling of CR. Gap-graded ARwet has higher or equal moisture resistance than similar Asphalt Concrete (AC). Open/Gap-graded ARwet has higher fatigue resistance than similar AC. Dense/Gap-graded ARwet has higher rutting resistance than similar AC. ARwet has a higher resistance to low temperature cracking than conventional AC. Gap-graded ARwet can reduce noise generation than conventional AC, but it is not always true.	Stiffness moduli of Gap-graded ARdry is less sensitive to high temperature than that of similar AC. Dense/Gap-graded ARdry has higher fatigue resistance than similar AC. Dense/Gap-graded ARdry has higher rutting resistance than similar AC. Dense/Gap-graded ARdry has similar rutting resistance than similar ARwet and Polymer Modified Asphalt (PMA) with SBS.	Dense-graded ARtb has the same or better moisture resistance as/than PMA. Dense-graded ARtb has the same fatigue resistance as PMA Dense-graded ARtb has similar resistance to low temperature cracking, compared to PMA Dense-graded ARtb has the same rutting resistance as PMA

Results of different field projects and laboratory tests showed that ARwet had promising mechanical performance. For example, the performance of ARwet on road E12 in northern parts of Sweden revealed that ARwet can fulfill the needs for long-life pavement with good flexibility for cold climate conditions [19]. Furthermore, laboratory tests showed that ARwet had lower modulus at lower temperatures and higher modulus at a higher temperature which are desired properties for resistance against low cracking and permanent deformation [20], [21]. Furthermore, the results of fatigue tests showed that ARwet had better fatigue cracking performance than conventional asphalt pavement [22]. ARwet had very good friction both immediately after compaction of pavement and after some years of traffic [23]. Besides positive results, results indicated that there was not any significant difference in stiffness and shear modulus, between the ARwet and conventional asphalt pavement [20]. Emissions of PM10 (Particulate Matter with aerodynamic diameters less than or equal to 10 µm [24]) due to studded tyre can be reduced on ARwet. However, the reduction of PM10 depends on type of asphalt pavement, e.g., open graded ARwet did not show any significant reduction in the emission of PM10, compared to conventional asphalt pavement [23].

According to the STA's tool for pavement management system, PMSv3, 154 km, out of 138,000 km public roads in Sweden, are constructed by AR (almost 0.1%). It should be noted that the available data in PMSv3 are only limited to public roads, so the AR pavement related to local roads, sidewalks and cycle roads as well as the roads possessed by municipalities are not reported. Fig. *1* shows the location of AR roads in Sweden. The north part of Sweden has 25 km of AR roads, the east part has 2 km, the Stockholm area has 19 km, the west part has 23 km and the south part has 85km of AR road.

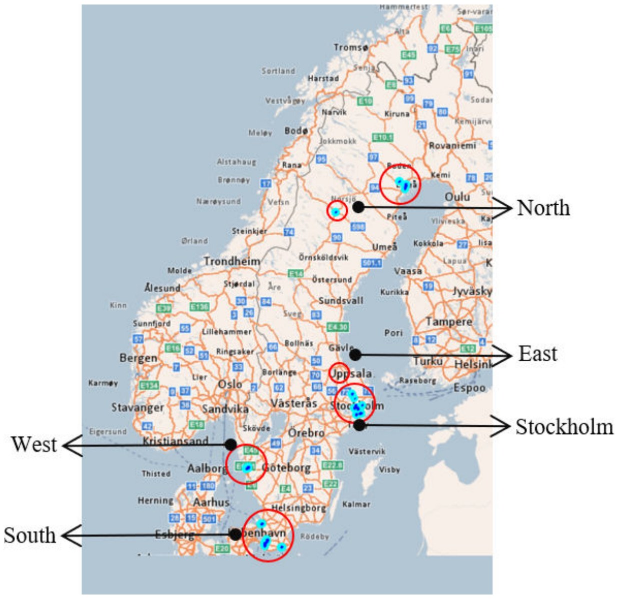


Fig. 1 Locations of asphalt rubber roads in Sweden, according to Swedish Transport Administration (svcPMSv3)

The production cost of ARwet in Sweden is about 20% to 30% higher than that of conventional asphalt pavement [25]. Hence, to compensate for the additional cost, it needs that ARwet roads provide at least 20% longer lifespan. The results of some laboratory tests and field studies indicated that the lifespan of AR pavements may be longer than that of conventional asphalt pavement [19], [22], [23]. Karri and Helwing [25] did a study to analyze the lifespan of AR in different climate zones in Sweden. The study was performed using available data from PMSv3. The data includes climate zone, asphalt type, average traffic volume, rutting depth and mean profile depth. The data were for five years from 2010 to 2014. The remaining lifespan of asphalt pavement was calculated based on the difference between the maximum acceptable rutting depth by SAT and the latest measured rutting depth in PMSv3, divided by the average variation of the rutting depth during five years. The quality of the data varied from climate zone to climate zone. In climate zone 1, where the data were more complete, the AR performed well, and the analysis concluded that in some cases the lifespan of AR was up to 29 years longer than the conventional asphalt pavement which had an average lifespan of 55 years. In climate zone 2, AR performed better than polymer-modified asphalt, on average had one year longer lifespan. For the other climate zones, it was not possible to make any conclusions, mostly due to a lack of data.

AR pavement, in comparison with conventional asphalt pavement, can have better mechanical performance, longer lifespan, lower maintenance cost and higher driving safety, however, implementation of AR poses some challenges, of which high temperature and health problems are the most critical for ARwet and low durability and very expensive costs for ARdry [26][27]. This study aims at analyzing drivers and barriers of the AR pavements in Sweden and also finding out the knowledge gap for future research on the field.

II. Method

The study is performed based on a literature review and by performing interviews with experts and relevant actors in the field. The interviews in this study are based on a semi-structured interview method. Semi-structured interviews are a common method used in qualitative research, in which the researcher prepares beforehand some questions and themes that the researcher wants to focus on. However, the interview is open, and the order of questions might change. Additionally, other questions might come up, to follow up on answers given by the interviewee [28].

The interviews, in this study, lasted between 60 min. to 90 min. and were performed digitally via Teams. In total 5 people were interviewed (Table 2), including asphalt researchers from VTI and asphalt experts from STA.

Table 2.
Relevant actors with experience in the field that were interviewed in this study

Date	Interviewee	Interviewee Competences
2021.09.22	Researcher 1 (R1)	Researcher with large experience (more than 30 years) in performing laboratory tests in multiple asphalt pavements. The researcher was involved in constructing and testing some rubber asphalt projects.
2021.09.28	Researcher 2 (R2)	Researcher with many years of experience in asphalt technology and design (more than 30 years). The researcher did the lab. tests and filed studies on rubber asphalt.
2021.09.28	Researcher 3 (R3)	Researcher with more than 10 years of experience in asphalt pavement design and test. The researcher did mechanical tests on rubber asphalt.
2021.10.06	STA asphalt expert 1 (STA1)	Expert/PhD on asphalt pavement technology and design. The expert worked for asphalt construction company for more than five years and during that period worked with the production of rubber asphalt, design and mechanical tests.
2021.10.10	STA asphalt expert 2 (STA2)	Expert on asphalt and binders. Large experience with chemical analysis on asphalt pavements, developing new asphalt pavements, and assessing asphalt product quality.

The interviews were divided into four main categories, namely: (i) technology, (ii) mechanical subjects, (iii) environmental impacts and (iv) others (including cost and policy). Each category was in turn divided into multiple subcategories, according to a literature review. Table 3 shows the main categories and correspondent subcategories covered in the interviews. Not all the subcategories were discussed in each interview. The interviews and discussion on each subcategory depend on the experience and knowledge of the interviewee.

Each interview was transcribed into text. The most important information was highlighted and summarized in a table, which included drivers and barriers in the rows, and the four main categories described in Table 3, in the columns. The knowledge collected in the interviews was coupled with information available in the literature to identify areas that need further research.

TABLE 3.
MAIN CATEGORIES AND CORRESPONDENT SUBCATEGORIES COVERED IN THE INTERVIEWS

1 - Technology	2 - Mechanical subjects
a) Fabrication of asphalt b) Construction of asphalt c) Technology & machinery	a) Type of AR pavement b) Type of binders used in AR in different climate c) Strength and weakness of AR pavement *d)* Rehabilitation and Maintenance
3 - Environmental impacts	**4 – Other (including Costs and Policy)**
a) Lifespan b) Health c) End of life of AR pavement d) Microplastic emissions e) Noise reduction f) Usage of raw materials g) Leaching	a) Construction cost b) Maintenance cost c) National and Regional legislation

III. RESULTS AND DISCUSSION

Table 4 shows the summary of the main drivers and barriers identified in the interviews as well as the comments from interviewees. Starting with the drivers, all the interviewees agree that the possibility of giving a new life to waste materials is a motivation to continue investigating the possibility of using AR in the future. Production of AR using available technology with small adjustments is another driver, referred to by most of the interviewees. For example, the technology which is needed to produce ARtb is very similar to the technology which is used to produce PMA. However, the cost of production of ARtb will be cheaper and the mechanical properties is comparable with PMA. Higher flexibility and better resistance against fatigue, thermal, and reflecting cracking are other drivers for AR, in comparison with conventional asphalt pavement, especially when AR is used as wearing layer on rigid pavement. The potential extended lifespan of AR can be counted as a driver as well, however, all interviewees said that there is no clear evidence that AR has an extended lifespan, and there is a need to further investigation on the functionality of AR over a longer lifespan than what has been done up to date. For a better understanding of the lifespan of AR, there is a need to update the previous studies [19], [22], [23], [25] by considering other important aspects such as annual average daily traffic, thermal and mechanical properties, type of asphalt, and production process.

There are also barriers to using AR identified in the interviews. One of the interviewees (STA2) pointed out that in general industries do not want to have the responsibility of taking care of wastes that are generated by other industries. This is directly coupled to another comment, referred to more than one time by the interviewees, that STA has some concerns about using waste as raw material. Therefore, in order to AR become more accepted by contractors and STA, there are needs of standardization to assure that CR, used to produce AR, does not expose risks to the environment and workers' heath. Production of ARwet needs higher temperature during production and construction as well as CR stocks to machinery during production. Furthermore, there is a lack of experience and design standards for the production and application of AR. Asphalt companies are not interested in the production of those types of asphalts that are not fully examined, and their application is assured. For example, the production of open-graded AR is difficult due to the high portion of air. In addition, one of the interviewees (R3) pointed out that the difference in stiffness and fatigue resistance between AR and conventional asphalt pavement is negligible. Another interviewee (R1) mentioned that the noise level on dense-graded AR and dense-graded conventional asphalt pavement is almost similar. It seems that the lack of a standard for the production of AR in Sweden leads to different results [20], [23]. Environmental barriers are also identified in the interviews. One of the interviewees (STA2) pointed out the need for further investigation on the risks of leaching of different substances that might be present in the AR. The interviewee highlighted that leaching studies should not only be limited to AR pavement but also leaching from all other asphalt types should be investigated. Gheni, A., et al. [29] assessed the leaching of using rubber aggregates in chip seal pavement under different pH conditions, in the US. The study concluded that for pH between 4 to 10, toxic heavy metals leached from the rubberized chip seal were below that of the EPA drinking water standards. The replication of similar studies in the Swedish context would allow for a better understating of the asphalt pavements leaching risk. The risk of microplastics or other AR components, spreading into nature, was also lifted during interviews. This is a subtopic with very little information in the literature. As leaching, the emission from microplastics can also be applied to AR pavements.

According to the interviews, the main challenge in the production and construction of AR pavements is work environment complaints, especially, due to terrible smell issues. The work environment and health problems were pointed out by two interviewees (STA1, STA2) as the main concern that made STA pause the use of AR. The interviewees point out that this problem must be overcome, if ARwet aims to be used more widely in the future. Nilsson T.P., et al. [30] investigated workers' exposure to particulates, polycyclic aromatic hydrocarbons (PAHs) and benzothiazole during the implementation of AR and conventional asphalt pavements in Sweden. The results showed that the respirable dust, total dust, particle number and mass, and total PAH concentrations for AR are the same as conventional asphalt. The levels of naphthalene, benzo(a)pyrene, and total particles were lower for both types compared with the Swedish occupational exposure limit. The study concludes that several air pollutants such as benzothiazole and PAHs are emitted into the air during asphalt work, but it is not evident if exposure to AR possesses a higher risk than exposure to conventional asphalt pavements in terms of asphalt worker exposure

It should be noted that ARdry does not have such problems but is very expensive to produce. Also, ARtb has not been yet produced in Sweden. Production of ARwet using Warm Mix Asphalt (WMA) method, which needs lower temperature during production, instead of Hot Mix Asphalt (HMA) method can be a solution to overcome the problem with ARwet.

Finally, the AR management after use was also referred to as a potential driver but also as a barrier. A driver, because old AR can be 100% recycled to produce new AR. However, it is important to understand the behavior of the new product, especially after the AR has been recycled many times. Additionally, one of the interviewees (R2) points out the possibility of recycling old AR pavement to produce conventional asphalt pavement and analyze how the recycled conventional asphalt pavement would behave. The information available in this field is very limited, and so it can be seen as a potential barrier The suggestion was to investigate how other countries, in which the use of AR is a common practice, are currently dealing with recycling and reusing of AR.

TABLE 4.
SUMMARY TABLE WITH MAIN DRIVERS AND BARRIERS IDENTIFIED IN THE INTERVIEWS

	Technology	Mechanical	Environmental	Other
Drivers	-By little adjustment, it will be possible to produce AR pavement using the existing technologies which are used to produce conventional asphalt (R1, STA1, STA2)	- AR is more flexible, compared to conventional asphalt (R1, R2, R3, STA1, STA2) -AR can reduce reflecting cracks when used as a wearing layer on a concrete pavement (R1, STA1) -AR has good resistance against temperature changes and thermal cracking (R1) - AR has probably a longer lifespan due to higher flexibility but it is not proven yet (R1, R2, R3, STA1) - AR has a better resistance against fatigue (STA1)	-Production of AR is a way to recycle waste tyre (R1, STA2) - AR can reduce noise (R1, R2, R3, STA1, STA2)	-Other countries produce and use AR (R1, R2, R39 -Interest from EOL tyres management (R1, STA2) -AR has no big mechanical difference, compared to PMA (STA2) -AR is more expensive than conventional asphalt but probably cheaper than PMA (STA2)
Barriers	-It is difficult to find the correct mix for AR (R1, R2, R3) -There is a lack of experience in applying AR (R1, R2, R3, STA1) - It needs to develop and produce different types of mixtures of AR (R1, R2, R3) - Higher temperatures are needed for the production and construction of AR (R1) - It needs more investment in AR technology before getting sure it works (R1, R2, R3) -AR gets stooked into the machinery due to the higher viscosity (STA2)	- More difficult to apply in open asphalt due to the high portion of air (R1) - There are no conclusive results if AR has better or worth mechanical properties than conventional asphalt pavement (R2, R3, STA1)	- Some complaints from workers about strong smoke and smell (R1, R2,R3, STA1, STA2) - A potential challenge is that AR can spread into nature (STA2) -Asphalt Industries do not want to be responsible for waste from tyre industry (STA2) -it needs that companies selling CR to assure that CR is safe to be used (STA2)	- Health problem is the main barrier for AR (STA1) - All actors in the asphalt industry do not have a full engagement to produce AR (R1) - Field projects of AR were at a small level and even some of them failed (R2, R3) -AR needs more bitumen and therefore might be more expensive (STA2) - There is no strict policy/standard to use environment-friendly asphalt (R2, R3)
Comments	- The failure cases for AR in Sweden were most probably due to lack of knowledge in the material mix, dimensioning and lack of application experience (R1, R2, R3) - First investment for the development of AR is high but in future the results will be promising (R1, R2, R3 - Temperature control during the construction of AR is not a big problem (STA1) - There is not any special standard to making and dimensioning AR (R2, R3, STA1)	- Fatigue and stiffness of AR are better than conventional asphalt but not that much (R2, R3) - AR has better or the same characteristics as conventional asphalt (R1, STA2) -AR can be produced in different types such as dense, stone mastic or open graded (R1, R2, R3, STA1) - AR can works for all traffic levels (R1, STA1) -AR can work well in cold climates (R1)	- There is not any special policy to produce and re-use AR (R2, R3, STA1) -EOL AR can be managed e.g. by mixing with newly produced AR (R1) - Research did not find any dangerous material in AR (STA1) -Emissions from AR pavements were not that different in comparison with conventional asphalt (R1, STA1) -It needs more research on how the AR can be recycled after use (STA1, STA2) - STA is not using AR because it is considered waste (R2, R3)	- To investigate the lifespan of AR, it needs field experiments (R2, R3, STA2) -AR needs optimization for cost and mechanical properties (STA1, STA2) - Rubber from tyre can be used in other technologies, even in a better way than asphalt (R2, R3) - PMB is easier to design and control than AR (STA2)

IV. CONCLUSION AND FUTURE SUGGESTIONS

This study aimed to compile previous knowledge on AR and to find drivers and barriers to the implementation of AR pavement in Sweden. The study was based on a literature review, complemented by interviews by experts.

There are generally three types of AR classification: ARdry, ARwet, and ARtb. ARdry was initially developed in Sweden and then exported to other countries. ARdry had good mechanical properties, however, it was very expensive in comparison with conventional asphalt pavements. The high cost of production made ARdry not be used widely in Sweden [27]. In 2006, ARwet was imported from the USA to Sweden and since then 154 km of roads have been constructed in Sweden using ARwet. The ARwet, especially dense graded, has better mechanical properties and a longer lifespan than conventional asphalt pavements. In the environmental perspective, extended lifespan comes with a potential reduction of resource use and thereby lower cost for construction and maintenance. However, complaints about terrible smoke and smell as well as concerns about workers' health problems made STA to pause production of ARwet in Sweden.

Considering new technologies in AR such as ARtb, better mechanical properties of AR than conventional asphalt pavement and lower cost of AR than PMA, it is necessary to do further research on the field. Some of the important knowledge gaps and interesting areas for future research and developments, found in this study, are:

- Health problems and how particle emissions during the implementation of AR can expose risks to workers.
- Environmental studies to analyzed the leaching of different types of substances into nature, including PHAs and heavy metals that might be present in the AR and CR.
- Comparison between ARtb and PMA – in terms of mechanical performance, but also costs and environmental benefits or drawbacks
- The cost of production of PMA and AR are about 25% more expensive than conventional asphalt pavement, so it is expected that PMA and AR should have at least 25% more lifespan than conventional asphalt pavement. It needs to analyze the lifespan of PMA and AR and compare it with the lifespan of conventional asphalt pavements.
- There is a lack of experience and design standards for the production and application of AR. Standardization of the design and production of AR is needed in the future.

Acknowledgment

This work was financially supported by Swedish National Road and Transport Research Institute. The authors are grateful to all interviewees for their valuable comments on the rubber asphalt.

References

[1] J. Patrício, Y. Andersson-Sköld, and M. Gustafsson, "End-of-life tyres applications: technologies and environmental impacts (VTI rapport 1100A)," Gotenburg, Sweden, 2021. [Online]. Available: http://vti.diva-portal.org/smash/get/diva2:1611409/FULLTEXT01.pdf.

[2] M. R. Pouranian and M. Shishehbor, "Sustainability assessment of green asphalt mixtures: A review," *Environ. - MDPI*, vol. 6, no. 6, 2019, doi: 10.3390/environments6060073.

[3] L. Brasileiro, F. Moreno-Navarro, R. Tauste-Martínez, J. Matos, and M. del C. Rubio-Gámez, "Reclaimed polymers as asphalt binder modifiers for more sustainable roads: A review," *Sustainability*, vol. 11, no. 3, p. 646, 2019.

[4] "Svensk Däckåtervinning," 2021. https://www.sdab.se/om-oss/statistik/ (accessed Dec. 17, 2021).

[5] Svensk Däckåtervinning, "Svensk Däckåtervinnings Årsrapport för 2019 tryckt och klar - Svensk Däckåtervinning," 2021. [Online]. Available: sdab.se.

[6] European Commission, "Circular economy action plan," 2020. doi: 10.2775/855540.

[7] M. Sienkiewicz, J. Kucinska-Lipka, H. Janik, and A. Balas, "Progress in used tyres management in the European Union: A review," *Waste Manag.*, vol. 32, no. 10, pp. 1742–1751, 2012, doi: 10.1016/j.wasman.2012.05.010.

[8] L. G. Picado-Santos, S. D. Capitão, and J. M. C. Neves, "Crumb rubber asphalt mixtures: A literature review," *Constr. Build. Mater.*, vol. 247, p. 118577, 2020, doi: 10.1016/j.conbuildmat.2020.118577.

[9] S. A. Alfayez, A. R. Suleiman, and M. L. Nehdi, "Recycling Tire Rubber Rubber in in Asphalt Asphalt Pavements : Pavements : State State of of Recycling the Art," *Sustainability*, vol. 12, no. 9076, pp. 2–15, 2020.

[10] R. Ghabchi, A. Arshadi, M. Zaman, and F. March, "Technical challenges of utilizing ground tire rubber in asphalt pavements in the United States," *Materials (Basel).*, vol. 14, no. 16, pp. 1–35, 2021, doi: 10.3390/ma14164482.

[11] S. Bressi, N. Fiorentini, J. Huang, and M. Losa, "Crumb rubber modifier in road asphalt pavements: State of the art and statistics," *Coatings*, vol. 9, no. 6, 2019, doi: 10.3390/COATINGS9060384.

[12] P. Rath, J. E. Love, W. G. Buttlar, and H. Reis, "Performance analysis of asphalt mixtures modified with ground tire rubber modifiers and recycled materials," *Sustain.*, vol. 11, no. 6, p. 1792, 2019, doi: 10.3390/su11061792.

[13] M. D. Nazzal, M. T. Iqbal, S. S. Kim, A. R. Abbas, M. Akentuna, and T. Quasem, "Evaluation of the long-term performance and life cycle costs of GTR asphalt pavements," *Constr. Build. Mater.*, vol. 114, pp. 261–268, 2016, doi: 10.1016/j.conbuildmat.2016.02.096.

[14] E. Y. Hajj, P. E. Sebaaly, E. Hitti, and C. Borroel, "Performance evaluation of terminal blend tire rubber HMA and WMA mixtures - Case studies," *Asph. Paving Technol. Assoc. Asph. Paving Technol. Tech. Sess.*, vol. 80, no. January 2018, pp. 665–696, 2011.

[15] H. T. Tai Nguyen and T. Nhan Tran, "Effects of crumb rubber content and curing time on the properties of asphalt concrete and stone mastic asphalt using dry process," *Int. J. Pavement Res. Technol.*, vol. 11, no. 3, pp. 236–244, 2018, doi: 10.1016/j.ijprt.2017.09.014.

[16] Federal Highway Administration (FHWA)., "FHWA-RD-97-148," 2016. [Online]. Available: https://www.fhwa.dot.gov/publications/research/infrastructure/structures/97148/st3.cfm.

[17] M. Frank and P. E. Rich, "Use of Tire Rubber in Asphalt Pavements in," North Dakota, USA, 1994.

[18] A. Shaker and S. Shaker, "Rubber Asphalt-asphalt paving mixed with ground tire rubber (in Swedish)," Jönköping University, 2013.

[19] A. Nordgren and T. Tykesson, "Dense graded asphalt rubber in cold climate conditions," *Asph. Rubber 2012*, 2012.

[20] A. Ahmed, H. Carlsson, and T. Lundberg, "Utvärdering av gummiasfalt – provväg E22 Mönsterås, Etapp 1," Linköping, 2019.

[21] T. Wang, F. Xiao, S. Amirkhanian, W. Huang, and M. Zheng, "A review on low temperature performances of rubberized asphalt materials," *Constr. Build. Mater.*, vol. 145, pp. 483–505, 2017, doi: 10.1016/j.conbuildmat.2017.04.031.

[22] S. F. Said, L. Viman, and T. Nordgren, "Provsträckor med gummiasflat ragn-sells infart vid granulat-anläggningen," 2014.

[23] L. Viman, "Gummiasfaltbeläggning Sammanställning av utförda mätningar och provningar," 2011.

[24] F. Oroumiyeh and Y. Zhu, "Brake and tire particles measured from on-road vehicles: Effects of vehicle mass and braking intensity," *Atmos. Environ. X*, vol. 12, p. 100121, 2021, doi: 10.1016/j.aeaoa.2021.100121.

[25] A. Karri and S. Hellwig, "Comparing rubber modified asphalt to conventional asphalt. Assessment of Trafikverket's road survey tool." 2015.

[26] A. Karagiannidis and T. Kasampalis, "Resource recovery from end-of-life tyres in Greece: A field survey, state-of-art and trends," *Waste Management and Research*, vol. 28, no. 6. SAGE Publications Sage UK: London, England, pp. 520–532, 2010, doi: 10.1177/0734242X09341073.

[27] T. Nordgren, "A Life Time of Experience with "Rubber Modification" of Asphalt Pavements In Sweden," 2018.

[28] J. Patricio, L. Axelsson, S. Blomé, and L. Rosado, "Enabling industrial symbiosis collaborations between SMEs from a regional perspective," *J. Clean. Prod.*, vol. 202, pp. 1120–1130, 2018, doi: 10.1016/j.jclepro.2018.07.230.

[29] A. Gheni, X. Liu, M. A. ElGawady, H. Shi, and J. Wang, "Leaching assessment of eco-friendly rubberized chip seal pavement," *Transp. Res. Rec.*, vol. 2672, no. 52, pp. 67–77, 2018.

[30] P. T. Nilsson *et al.*, “Emissions into the air from bitumen and rubber bitumen - Implications for asphalt workers’ exposure,” *Ann. Work Expo. Heal.*, vol. 62, no. 7, pp. 828–839, 2018, doi: 10.1093/annweh/wxy053.

Sign Projection Lamp – A Design Approach

Anju Kumari, Akhilesh V Madathil, Lalit R Ahuja, Vaibhav Baranwal

Reasearch & Development – Autotmotive Lighting, Varroc Engineering Limited, Pune, 411033, India

Abstract — We propose a design method for the sign projection lamp based on Koehler Illumination. It is a combination of condenser optics and projection optics. Both the condenser optics and projection optics are the combination of multiple aspheric lens surfaces. The image film needs to be placed in between the condenser system and projection system. The Condenser optics provides uniform collimated rays on the image film collected from the light source and projection optics together provides the images on the different planes with minimal aberrations.

Keywords—Symbols, Projection, LED, road safety, Koehler Illumination

I. INTRODUCTION

Lighting industry has followed the same trajectory in terms of advancements as the motorization did, in automotive sector in last few decades. Automotive lighting industry always try hard to improve night drive situations and to reduce the road traffic deaths. The evolving technology for enhancing the road safety such as self-driving cars and adaptive automotive lighting systems help reducing the traffic accidents but not offering much in pedestrians death rates. According to WHO, the pedestrians, cyclists and motorcyclists are the most vulnerable road users (VRUs) which comprise half of the global death [1]. Communication among vehicles and pedestrians can play a vital role in reducing the death rates across the globe. Introducing a technology which enables drivers to communicate with other road users through projection of symbols on road can act as an alarm which can contribute to reduction of road fatalities.[2]

Talking about the technologies calling for high cost, such as DLP (Digital Light Processing) Projectors using DMDs (Digital Micromirror Devices) by Texas Instruments and Laser scanning projection system [2], can be used to project different images on road to establish communication among vehicles and pedestrians but keeping cost in mind, there is a need to develop a projection system which is favorable to adapt anytime and, in any vehicle, [3][4]. More recent technologies include projector with array LED matrix light source and laser projector systems. [5][6]

Although it is still an issue to introduce the projection of symbols and signs into the UNECE (United Nations Economic Commission for Europe) R48 and R149, and there is difference in opinions between some related parties such as GRE (Working party on Lighting and Light Signalling) and experts from GTB (The International Automotive Lighting and Light Signalling Expert group), the intention should be clear for making the safety as priority and implementation of these under well-defined conditions. While the debate shall be in commotion among them, it is important for us as lighting manufacturer to be ready with solutions that can enhance the overall pedestrian safety without adding any distraction for others.

A design is proposed here which has simple mechanism but out-standing performance in low cost. Using a projection system, we can project symbols to enhance road safety and logos for styling purpose.

II. DESIGN

A. Approach

This section describes the design approach for image projection system. An optical system that contains combination of lenses with specific functions is used for projecting the image of slide/film at a required distance on the road called a projection system. It comprises a light source, condenser lens, film, and projection lens system. Light guide could also be used as condenser optics, but it may affect the compactness of the design [2]. Laser based projection systems are developed where laser is used as a source [5]. An LED (Light Emitting Diode) is chosen here as the light source.

LEDs are durable and long lasting with less power consumption [5]. A suitable LED can be employed as per the requirement of color and brightness of the image. Divergent light from the LED will be collected by the condenser lens and illuminates the film completely. The film/slide is equipped with required symbol imprinted on it with transparency. Projection lens forms the image of the film on the screen at required distance as it finds the object to be the film.

The LED is a Lambertian light source which follows Lambert’s cosine law [8]. It states that illumination on any surface at any point E, is given by the cosine of angle between the normal to that surface and line of flux. LEDs emit light in different angles covering entire semi hemisphere. For collecting the maximum light from the

LED to make an efficient imaging system, we need a condenser optics which not only collects the major amount of light from the light but also provides uniform illumination to the film. There can be many different optics for collecting and converging the light beam including spheric/aspheric lenses, collimators etc., it is important to consider the Numerical aperture (NA) of the condenser optics system to accept the light.

The condenser optics will consider the source as the object and will image it in space may be at infinity. The film needs to be placed after the condenser optics which will be object to the projection optics.

The rays deviate from ideal behavior and fails to converge at a point results in the aberrations. There are two primary causes of non-ideal lens action: Geometrical or Spherical aberrations are related to the spherical nature of the lens and approximations used to obtain the Gaussian lens equation; and Chromatic aberrations, which arise from variations in the refractive indices of the wide range of frequencies found in visible light [12]. Aberrations in the image can be controlled by the optimizing the overall projection system, these include materials, radius of curvatures and/or thicknesses of the components.

Magnification is another important factor in projection systems. It is the size of an image relative to the size of object producing it [13]. Magnification can be introduced according to the requirements, by the projection system.

Koehler illumination technique is used in this projection system. It is used to ensure that the final image doesn't contain any image of source. Koehler illumination technique was developed by August Koehler in 1893. If the film plane has an image of source on it, then the final image will also consist of the image of source along with image of film. Therefore, Koehler illumination is of great importance in imaging systems.

Key points of Koehler's illumination [6][7]:

I. Koehler's illumination is used when the source is non-uniform like tungsten filament, LED etc. It results in an evenly illuminated image.

II. Koehler's illumination is characterized by forming the image of the source at the entrance pupil of the projection optics through the film.

III. The film consisting the symbol must be placed close to the condenser optics to illuminate it completely. Placing the film at the exit pupil of condenser will illuminate it completely and uniformly.

IV. The image of the film is produced by the projection optics on the screen.

V. There should not be any image of source at the film plane otherwise, the final image of the film will contain the image of source in it.

VI. The size of image of source should be equal to the entrance pupil of the projection optics.

B. Design Aspects

In this paper, we decided to design an image projection system with depth ~ 55 mm, and diameter within 30 mm. This size has been decided considering the RFQ requirement for such kind of lamp and space available to place it near stand area or in headlamp in two-wheeler. In this section, the description of components will be provided. Zemax is used to design and optimize imaging systems.

1. Source

LED is a Lambertian source and divergent in nature. The aim should be to cover the maximum flux from the LED and directing it into the system. The condenser system is used to serve this purpose, which is discussed in the next part. The closer the LED is placed to the condenser system; more will be the input energy into the system, and we obtain a more efficient system. In this design, LED is kept at 4.139mm from the condenser optics. Decreasing the distance between the two can further increase the efficiency but it becomes difficult to design a condenser system of such small focal length. The half cone angle of acceptance by the first lens is 59.99 degrees. The efficiency comes out to be 86%. Now, the Numerical aperture for the condenser lens must be known for defining light collection from LED. The decision depends on the divergence angle of LED, usually tend to be 180 °.

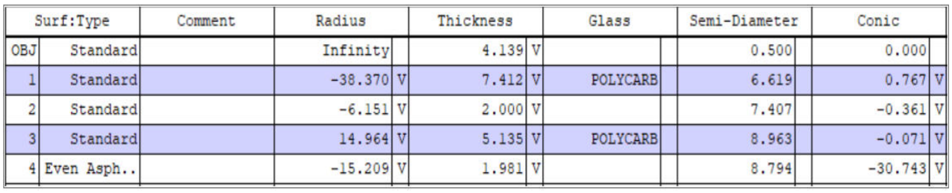
$$\mathrm{NA}=n\ sin\theta=1.0\times\sin(59.99°)=0.866 \qquad (1)$$

Θ is the maximal half cone angle that can enter/exit the lens.

n is the refractive index of working medium of lens.

2. Condenser Optics:

Considering the input energy to the system, a condenser optics having NA of 0.866 is designed. Two positively powered lenses have been used. Figure-1 is representing the condenser system. One is to collect the light from the LED and another one to converge and focus it onto the entrance pupil of the projection system. Different kind of optics can be used



Surf:Type		Comment	Radius	Thickness	Glass	Semi-Diameter	Conic
OBJ	Standard		Infinity	4.139 V		0.500	0.000
1	Standard		-38.370 V	7.412 V	POLYCARB	6.619	0.767 V
2	Standard		-6.151 V	2.000 V		7.407	-0.361 V
3	Standard		14.964 V	5.135 V	POLYCARB	8.963	-0.071 V
4	Even Asph..		-15.209 V	1.981 V		8.794	-30.743 V

Table-1-condenser system

to condense the light from the LED including collimators, freeform lenses, light guide etc. This optics consisting of two lenses is preferred to make a condenser system considering the space constraints and manufacturing.

Table-1 is describing the parameters of the condenser system for each surface. The lenses have both spherical and aspherical curvatures. The image of source should have magnification comparable to the size of the entrance pupil of the projection optics to completely illuminate it or to render the entire light from source.

The first element of the condenser system has thickness of 7.412mm with diameter 13.238 mm and the second has thickness of 5.135 mm with diameter 17.926 mm. The airgap between the two is of 2 mm. Glass or plastic optical materials can be used depending upon the application. In this design, lens elements of condenser optics are made up of PC (Polycarbonate) with a refractive index of 1.586.

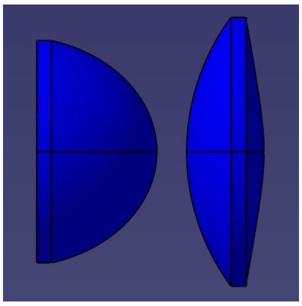


Fig-1-Condenser system

3. Film

The film is the object for Projection lens. It is 0.5 mm thick and contains the symbol on it. The dimensions of the symbol will lead to the dimension of image with magnification of the projection lens. A symmetrical symbol is used as shown in Fig-2, to analyse the higher field rays. The symbol will be transparent and rest of the area on the film will be opaque.

It can be made up of materials viz. PC, PMMA etc. for Plastics. Here PMMA is used.

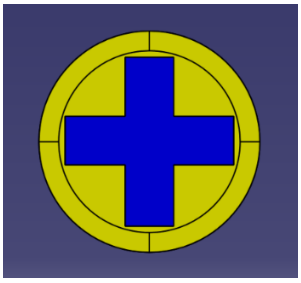


Fig-2-film

4. Projection Optics

This is the heart of the system that projects the image of the symbol on the road with required specifications. The projection optics decides the magnification provided by the system. Figure-3 is showing the projection system.

This design can be exploded into four individual lens elements. i.e., two positive lens elements and two negative lens elements. These four lenses together form a positively powered system, and this property will result in an inverted image of object on the screen.

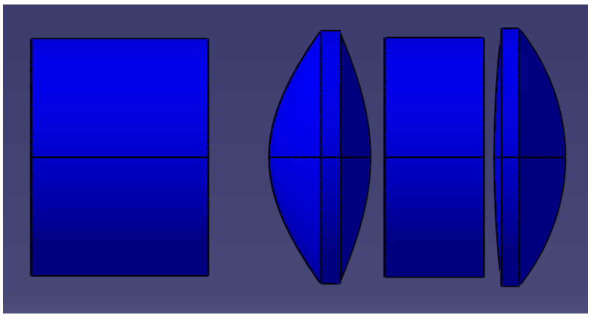
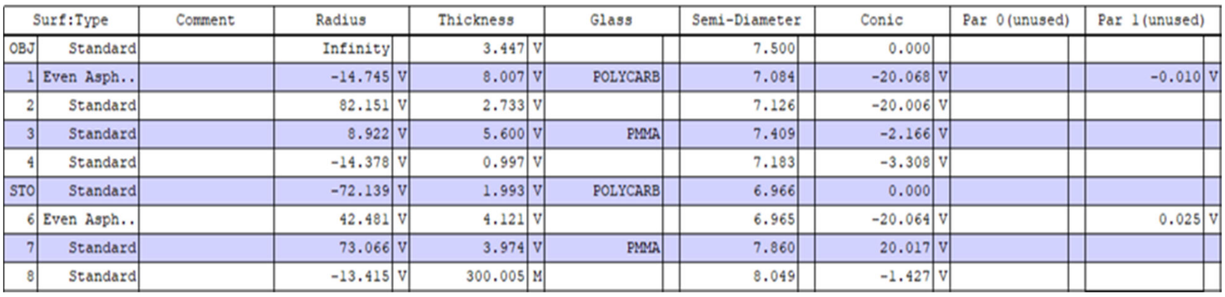


Fig-3-projection system



Surf:Type		Comment	Radius	Thickness	Glass	Semi-Diameter	Conic	Par 0(unused)	Par 1(unused)
OBJ	Standard		Infinity	3.447 V		7.500	0.000		
1	Even Asph..		-14.745 V	8.007 V	POLYCARB	7.084	-20.068 V		-0.010 V
2	Standard		82.151 V	2.733 V		7.126	-20.006 V		
3	Standard		8.922 V	5.600 V	PMMA	7.409	-2.166 V		
4	Standard		-14.378 V	0.997 V		7.183	-3.308 V		
STO	Standard		-72.139 V	1.993 V	POLYCARB	6.966	0.000		
6	Even Asph..		42.481 V	4.121 V		6.965	-20.064 V		0.025 V
7	Standard		73.066 V	3.974 V	PMMA	7.860	20.017 V		
8	Standard		-13.415 V	300.005 M		8.049	-1.427 V		

Table-2-projection system

PMMA is used for the positive lenses and PC is used for negative lenses. The projection system has an effective focal length of 18.6187 mm.

Table-2 is showing the complete information about the parameters used in designing the projection system including radius of curvatures for each surface, thicknesses and air gaps between each surface, materials, dimeters, conic constants etc. The thicknesses for each lens element are 8.007 mm, 5.6 mm, 1.993 mm and 3.974 mm. The air gaps between lens elements are 2.728 mm, 2.209 mm and 4.144 mm.

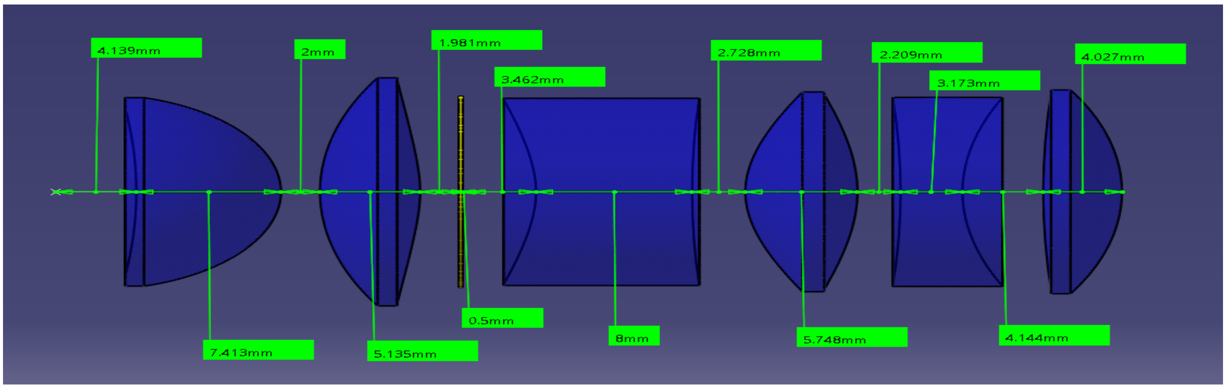


Fig – 4 – Complete Lamp System

Lens semi-diameters for each element are 7.084 mm, 7.409 mm, 6.966 mm and 7.860 mm

5. Complete System

After merging the projection optics and the condenser optics, complete system is obtained. Fig-4 is showing the complete system where source (LED) position is shown with a point. We have total of six lens elements. The depth of the lamp is 55 mm including the LED distance and diameter of the lamp is 26 mm.

The propagation direction is from left to right. The object is kept inverted to produce an erect image following the geometrical optics calculations.

To obtain an image at a location far by 300 mm, object/film is placed at 3.447 mm from first lens surface of the projection system. The projection lens brings a magnification of 14.89x to the object, keeping object size 11.95 mm, image of 175 mm is obtained.

Projection system controls the aberrations for this lamp. The aberrations are the imperfections in the image. The basic seidel aberrations are spherical, coma, distortion, field curvature, astigmatism and chromatic aberration due to different wavelengths.

Considering the combination of different materials like crown and flint can reduce the chromatic aberration along with few others. PC is flint and PMMA is crown in nature according to abbe's number. Aspheric lens surfaces are used to reduce spherical aberration and keeping small the aperture diameter can help minimising the overall aberrations. Sometimes, we need to introduce negative aberrations to compensate positive aberrations and vice versa.

It’s difficult to eliminate aberrations completely, so knowledge of aberration tolerance is required to design a system as per requirements.

III. SIMULATIONS RESULTS

The Varroc ray tracer is used to simulate and obtain the result. Varroc ray tracer is an in-built tool in Catia software in-housed by Varroc.

Fig-5 is showing the raytracing for this system and the angular spread of rays. The four black lines are the screens at distances 200 mm, 300 mm, 400 mm and 500 mm from the lamp. According to the distance of image from the principal plane of projection optics, object/film distance can be calculated using the lens formula. Increasing the screen distance from the lamp results in decreased intensity values, according to the inverse square law for intensity. Also, the size of the image will keep on increasing following the magnification equation.

Though any symbol can be projected with this lamp, a symmetrical symbol has been chosen here to analyse the higher field rays. Fig-6 is showing the simulation results at different screen distances from the lamp.

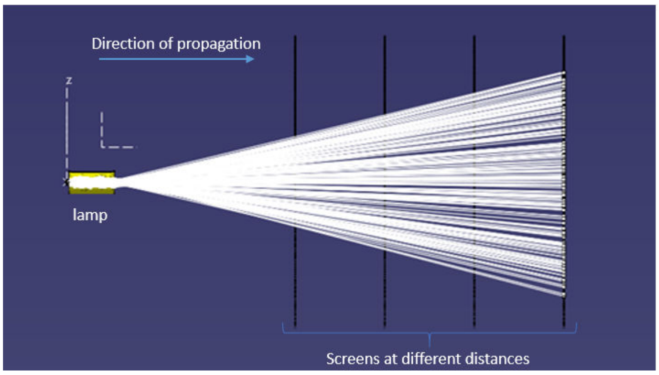
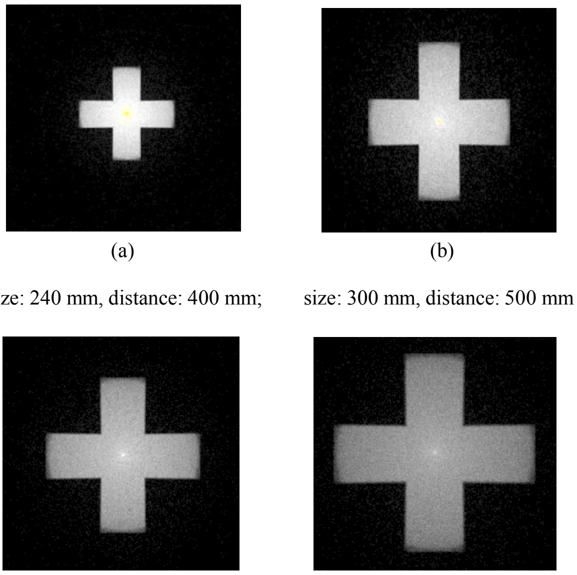


Fig-5- Raytracing

The (a), (b), (c), (d) are images at screen distances 200 mm, 300 mm, 400 mm, 500 mm respectively. The sizes are mentioned in fig 6. The screens are of size $400\ mm \times 400\ mm$. The input luminous flux is 1 lumen. The luminance at screen distances 200 mm, 300 mm, 400 mm, 500 mm are 21.2 nits, 10.4 nits, 6.18 nits and 4.09 nits respectively and the last three are normalized with respect to the first one. Images of uniform brightness are obtained.

size: 118 mm, distance: 200 mm; size: 175 mm, distance: 300 mm



(a) (b)

size: 240 mm, distance: 400 mm; size: 300 mm, distance: 500 mm

(c) (d)

Fig-6- Images at different screen distances

IV. DISCUSSIONS AND FUTURE DEVELOPMENTS

The present study is done with an intent to produce a white color image, but different color images can be produced using an appropriate light source or using different color filters fused in the film itself. The system is designed to focus on a certain distance with narrow depth of field, but there is possibility to design a system with wide depth of field to get a focused image across a wide distance. Designing a multifocal system offers the possibility to place the lamp anywhere in the vehicle to serve different purposes because it screen distance can be varied but still we will get a focused image. This lamp can be used to project any symbol. Fig-7 is showing images of a left turn and a right turn sign, giving an example of images that can be displayed on road using this projector.

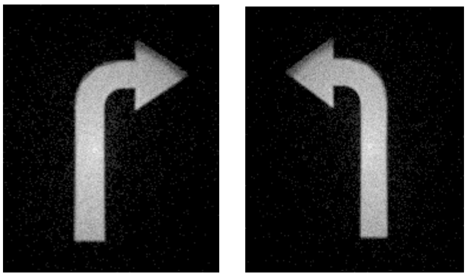


Fig-7-turn signs

V. REFERENCES

1. https://www.who.int › GSRRS2018_Summary_EN

2. Masashige Suwa, Masatoshi Nishimura and Reiko Sakata," LED Projection Module Enables a Vehicle to Communicate with Pedestrians and Other Vehicles" IEEE International Conference on Consumer Electronics (ICCE)-2017, 2158-4001 (DOI-10.1109/ICCE.2017.7889220)
3. https://www.ti.com- dlp-chip
4. Maoshegn Hou, Zhengxue Shi, Jiqiang Liu, Yuqing Chen, and Tianxiong Li, "Development of a laser scanning projection system with a dual-diameter fitting method and particle swarm optimization" Vol. 60, Issue 5, pp. 1250-1259 (2021)
5. https://www.sharpnecdisplays.eu/p/laser/en/technologies.xhtml
6. https://www.osram.com/os/applications/automotive-applications/eviyos-digital-light.jsp
7. http://www.ledlights.org
8. https://en-academic.com/dic.nsf/enwiki/26168
9. Angelo V. Arecchi, Tahar Messadi, R. John Koshel," Field Guide to Illumination" www.spiedigitallibrary.org/ebooks/FG/Field-Guide-to-Illumination/ ISBN: 9780819481221 (2017)
10. https://www.zeiss.com- Koehler illumination
11. https://www.olympus-lifescience.com/en/microscope-resource/primer/anatomy/aberrations/
12. https://www.britannica.com/technology/magnification
13. Joseph M. Geary," Introduction to lens design- with practical ZEMAX examples" ISBN-10: 0943396751 (2002)

Efficient utilization of negative half wave of Regulator Rectifier Output to drive Class D LED Headlamp

1st LALIT AHUJA
2-W Lighting
Varroc Engineering Ltd.
Pune, India
lalit.ahuja@varroc.com

2ndYashas Shetty
2-W Lighting
Varroc Engineering Ltd.
Pune, India
yashas.shetty@varroc.com

3rd Nancy Das
2-W Lighting
Varroc Engineering Ltd.
Pune, India
nancy.das@varroc.com

Abstract— **LED lighting has been increasingly adopted for vehicles in both domestic and foreign automotive markets. Although this miniaturized technology gives the best light output, low energy consumption and cost-efficient solutions for driving the same is the need of the hour.**

A novel vehicle headlamp driver for two-wheeler Class D LED headlamp is presented here. In this paper, unlike usual LED headlamps which is driven by battery, Regulator and Rectifier (RR) driven, low cost and highly efficient LED Driver Module (LDM) is proposed.

Our system uses negative half wave rectified DC output from RR to provide a constant light output at all RPM values of vehicle. With a negative rectified DC output of RR, we have an advantage of pulsating DC input which periodically goes to zero thus help us to generate a constant DC output equivalent to required LED load and with change in RPM additional active thermal bypass circuit help us to maintain the efficiency and thermal rise.

Keywords— Class D LED headlamp, Regulator and Rectifier (RR), Pulsating DC, Low cost and highly efficient, LED Driver Module (LDM).

I. Introduction

In this paper, we present methodology for driving highest class two-wheeler headlamp with Regulator and Rectifier (RR) output.

The positive half of the magneto output is regulated and used to charge batteries used for various peripherals. While, conventionally, the negative half was used for operating bulb based exterior lamps. But with advancement in LED based headlamps which are driven by battery, this negative half pulse remained unused in most of the vehicles.

Hence, the aim is to utilize the unused negative pulsating DC output of RR, so as to optimize utilization of RR output power and provide a cost-efficient solution as compared to costly DC-DC drivers. This paper presents a novel idea to use the negative half wave output of the RR along with a linear constant current driver with significantly higher efficiency to drive up to 30W LED loads. Although the RR output has varied frequency and duty cycle at different engine RPMs, the driver is designed as such that it provides constant current to LEDs at all engine RPMs with minimal ripple.
A switching regulator works by taking small chunks of energy, bit by bit, from the input voltage source, and moving them to the output. This is accomplished with the help of an electrical switch along with magnetic components and other passive components which result in bulky and costly components. But with linear regulators, we're eliminating bulky components and improving the form factor. Hence, the proposed solution is both cost efficient and compact.

Presently, output ripple free amplitude drivers with fewer components and less complexity are limited to lower power LED Lamps. The focus of current high-efficiency research is often on high LED power applications. This paper presents a method of driving LED load at both High Beam and Low Beam using the negative half wave rectified pulsating DC from RR with minimum number of components, maintaining high efficiency within the thermal limitations. The project is undertaken with **Varroc Engineering Ltd. – 2-wheeler Lighting Electronics Team.**

Usually using linear regulator leads to poor thermal performance, as linear regulator works by taking the difference between the input and output voltages, and just burning it up as waste heat. The larger the difference between the input and output voltage, the more heat is produced. In most cases, a linear regulator wastes more power stepping down the voltage than it actually ends up delivering to the target device. With typical efficiencies of 40%, and reaching as low as 14%, linear voltage regulation generates a lot of waste heat which must be dissipated with bulky and expensive heat sinks. This also means reduced battery life. Linear regulators are great for powering very low powered devices. They are easy to use and cheap, and therefore are very popular. However, due to the way they work, they are extremely inefficient.

But with the input being negative half wave rectified pulsating DC from RR, this efficiency can be improved. As the input periodically goes to zero, this helps us to generate a constant DC output equivalent to LED load, thus minimising the voltage drop on the linear regulator. As the drop on the linear regulator is minimal, losses are significantly reduced and efficiency as high as 75% is achieved. With change in RPM, this DC voltage increases which can be managed by active thermal bypass circuitry, thus resulting in better thermal performance. Hence, use of bulky and expensive heat sinks can be avoided. As the battery input is not connected to the lamp now, battery life is also unaffected.

The constant DC output used to drive the linear regulator and the LEDs is obtained without using any magnetic components and freewheeling diodes. A simple switch and capacitor are arranged in such a way that the voltage across the capacitor is DC with some ripple. The switch ON time and OFF time are dependent on the input voltage amplitude.

Conventionally, with battery input, driving large array of LEDs requires voltage to be stepped up, hence switching topologies are essential. But using negative half wave sine input, with peak reaching as high as 35V, switching regulators are no longer required.

For low load voltage applications, stepdown switching regulators are must. But with input periodically going to zero, setting required stepped down voltage can be obtained without switching regulators.

II. Challenges with negative half wave sine input

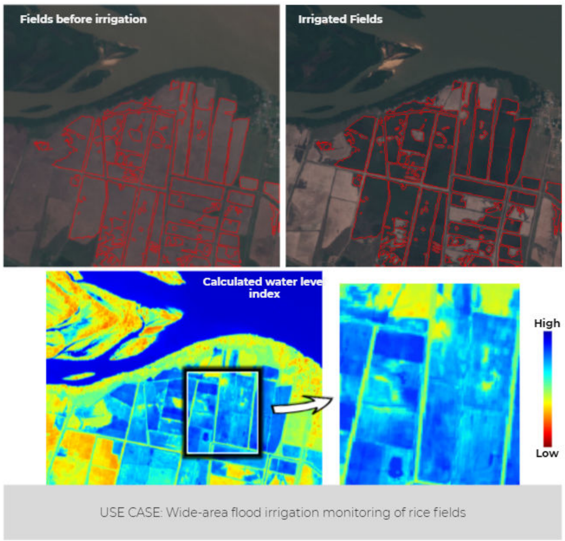
- The current drivers used in LED applications are driven by battery which has nearly constant voltage. Whereas the amplitude of negative half cycle of the RR varies with RPM.
- With input voltage periodically going to 0 Volts, maintaining constant voltage at input such that it is sufficient to drive the LED load with minimum losses of LED driver circuitry.
- The voltage threshold set needs to be such that it helps to maintain constant voltage across linear regulator which is equivalent to drive the LEDs.
- As the output of voltage limiter circuitry is not constant DC but rather rippled DC, additional circuitry for thermal management is required to minimize load on linear regulator

State of Art Solutions Based on Hyperspectral Satellite Imagery

Eldrige de Melo – eldrige.melo@satellogic.com

Introduction: Detecting land and water features through the use of Hyperspectral Satellite Imagery. For the past decade, the world has been bracing for different global challenges such as growing populations and global warming. In order to tackle these challenges and their consequences, hyperspectral satellite imagery can be used as a cost-effective operational data for timely decision making. Satellogic has developed tools and techniques that extract valuable data from hyperspectral satellite imagery that can assist to mitigate the effects of climate change and over population that are estimated to have a higher impact in developing regions.

One of the main consequences of over population and global warming is food shortage, this research provides solutions based on the use of Satellogic's capacity to acquire 25 cm Hyperspectral images on a daily basis globally. This capacity allowed this research project to create state of the art solutions such as irrigation monitoring that provides early detection of irregular irrigation patterns allowing for quick response and increased productivity.



USE CASE: Wide-area flood irrigation monitoring of rice fields

More advanced farming techniques such as crop estimation through campaign management, species selection and dynamic yield selection.

Risk-Based Prioritization Approach to Quality Control Inspection Activities On Structural Works for High-Rise Buildings in Makati City

P.G. Aguada, K.D.P. Cinco, R.R.R. Reyes

***Abstract* -** The purpose of these inspections is to determine the status of the construction project and to decide what steps should be taken to approve, reject, or modify such part of that recognition. The inspection activities necessitate that these inspectors be on the move to record the possible concerns in the future relating to a specific project problem.

In the industry, each of the companies have corresponding standards being set and followed upon the different work scopes involved in operations. Criteria, protocols, and guidelines are considered as a critical guarding factors, specifically in assessing quality assurance inspection on structural works, for this scope which involves high-risk activities.

A good quality control inspection has a major impact on both businesses in architecture and engineering. And a company constantly committed to improved quality has the higher chance of acquiring zero defects and fatalities, appearing to a competitive rank of quality performance, risk-management, and overall accomplishments.

I. INTRODUCTION

Construction inspection plays a very important role in the industry as it is an incentive for cost reduction, efficiency improvement, and a significant competitive advantage. It is also an important part of the quality control program for ensuring quality and long-term success of a project. The purpose of these inspections is to determine the status of the construction project and to decide what steps should be taken to approve, reject, or modify such part of that recognition. To carry out the inspection, one or more quality inspectors work individually or by forming an inspector team. The inspection activities necessitate that these inspectors be on the move to record the possible concerns in the future relating to a specific project problem.

In the industry, each of the companies have corresponding standards being set and followed upon the different work scopes involved in operations. Criteria, protocols, and guidelines are considered as a critical guarding factors, specifically in assessing quality assurance inspection on structural works, for this scope which involves high-risk activities. A good quality control inspection has a major impact on both businesses in architecture and engineering. And a company constantly committed to improved quality has the higher chance of acquiring zero defects and fatalities, appearing to a competitive rank of quality performance, risk-management, and overall accomplishments.

Restie Ross Reyes is with the Mapua University, Philippines (e-mail: restyross.reyes123@gmail.com).

II. METHODOLOGYY

This chapter introduces the methods used in this study which are the following: (1) carrying out an extensive literature review and conducting an initial interview with Quality Control Inspectors and Structural Field Engineers to identify and review several quality control inspections in structural works for high-rise buildings, (2) designing a questionnaire to collect data to assess the frequency and severity of identified risks, and (3) performing reliability analysis and ranking analysis for these collected data.

Primary Data Collection

The identification of risks in QC inspection activities in structural works is based on extensive literature review. An additional interview with QC Inspectors and Structural Engineers is to be conducted by the researchers online to review these inspection activities and to evaluate initially its risks.

Secondary Data Collection

After identifying risks in QC inspection activities in structural works, a survey questionnaire is designed for respondents from the Construction industry. These respondents have already had, and majority are still working on construction field specifically in high-rise vertical buildings in Makati City. The questionnaire that the researchers prepared are divided into two sections: survey questionnaire enlisting four sections; respondents' profile, level of awareness and familiarity regarding the quality control structural work inspections, level of agreement in their working experiences and based on their background focusing on structural works, and risk-based strategies as significant options to increase quality and safety on construction site. All these sections entailed in the survey questionnaire shall be collected and analyzed through the Analysis of Variance or ANOVA Statistical Treatment. On the other hand, for the interview segment, the researchers made use of the online platform to provide online questionnaire and send them to the chosen respondents, particularly the Quality Control Inspectors and Structural Engineers, via e-mail in the form of online document for approval of their available schedule and will then be proceeding to zoom meetings for the main part of the interview which then consists of interactive question and answer portion between researchers and respondents.

researchers to identify what should be the outline flow and categories to be considered in the survey questionnaire for the respondents. This will also help the researchers in coming up what are the improvements or revisions that a certain construction firm should be taken in consideration to avoid work stoppage and deficiencies with regards to project quality and decrease in employee population because of lack of safety and security which causes also their work productivity to decrease during the pandemic crisis.

Data Analysis

Through reliability analysis, Cronbach's alpha (Equation 1) is used to measure the reliability of the questionnaire where:

n is the number of risk factors

σ_x^2 is the variance of the observed risks,

σ^2 is the variance of component i for the

current sample of respondents?

For data to be internally consistent, the value of the alpha must be at least 0.60.

$$\alpha = \frac{n}{n\%\&}\%1 - \frac{\text{Zi"1}^{n}\ \sigma^{\&}_{Yi}}{\sigma^{\&}_{x}}\ |$$

After this, ranking analysis is utilized to determine the frequency and severity indices of the identified and assessed risks from the survey data. The mean value of the rating for the importance of risk is considered as the importance index and the mean value of the rating of the frequency is considered as the frequency index. It is used to rank importance indices of all risks and frequency of occurrence of all risk factors, respectively (Polat et al, 2017). To measure its overall ranking, the importance index (Equation 2) and the frequency index (Equation 3) are multiplied to get the severity index (Equation 4).

$$Importance\ inde\ = \frac{)n1*+n\&*3n(*2n)*n*}{)(n_1*n_\&*n_(*n_)*n_*)}$$

Where $n_\&$ is the number of responses for Very High, n2is the number of responses for High, n3is the number of responses for Moderate, n+is the number of responses for Low, and n) for the number of responses for Very Low.

$$Frequency\ inde\ = \frac{)n1*+n\&*3n(*2n)*n1}{)(n_1*n_\&*n_(*n_)*n_*)}$$

where n&is the number of responses for Always, n2is the number of responses for Almost Always, n3is the number of responses for Often, n+is the number of responses for Sometimes, and n) for the number of responses for Rarely.

$Severity\ index = Importance\ index \times Frequency\ index$

Research Setting

In line with the occurrence of Covid-19, where face to face exertions have been limited, the research will be conducted by an online survey. The online survey will be answered by Quality Control Inspectors and Structural Engineers who face risks daily on their structural works on High-Rise Buildings in Makati.

Respondents of the study

The respondents for this research for the survey gathering procedure will be thirty (30) Quality Control Inspectors, Safety Officers, and Structural Field Engineers of three (3) construction companies, who face risks in their works on High-Rise Buildings. The companies where the participants work must be based in the Makati area, since this is the setting of the study.

Data Gathering Procedures

During the time of the research, Covid-19 is still affecting the Philippines and transactions were set to be limited. Following the health and safety guidelines, the researchers gathered data:

1. A series of questions will be developed in Google docs. Google docs will serve as the questionnaire for the respondents.
2. The researchers will explain to the correspondents the purpose of the survey.
3. The target respondents for the research are quality control inspectors and structural engineers.
4. The survey will be about 10-15 minutes to accomplish
5. The responses will be analyzed and discussed by the research and come up with a conclusion.'

III. STATISTICAL TREATMENT

After performing the instruments to gather data, the researchers will analyze the data from the respondents through the outcome from the chosen respective statistical formula, Analysis of Variance. After analyzing the data presented, the researcher will formulate a conclusion based on their interpretation of the data.

$$F = \frac{\Sigma n_j(\bar{X}_j - \bar{X})^2 / (k-1)}{\Sigma\Sigma(X - \bar{X}_j)^2 / (N-k)}$$

Equation: Analysis of Variance

Analysis of Variance Test is used to determine the difference between two or more means of a population. Anova test used the variation, variation of between mean and the variance of the samples. Using the P-value F-Critical, F-value, generated from the of the analysis of variance test will verify the significance of the test considering significance level of 0.05.

IV. DEMOGRAPHIC PROFILE

1. The first section of the questionnaire is the profile of the respondents in terms of age, gender, years of experience, current designation, years of experience, and type of organization.

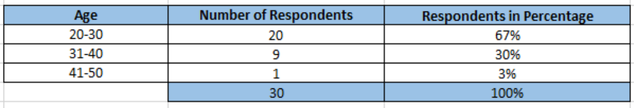
2. The second section of the questionnaire is the level of awareness of the respondents regarding structural quality inspection and structural work sequences and procedures.

3. The third section of the questionnaire is the level of agreement to working on construction sites in accordance with their respective profile and background about structural works in terms of quality control standards.

4. The fourth section of the questionnaire is the determination of action plans and risk-based strategies to increase safety and project quality and achieving minimal conflicts on quality control inspection on structural works.

The Profile of the Respondents in Terms of Age, Gender, Years of Experience, Current Designation of Construction Site Employees, and Type of Organization

Respondents in terms of Age



Age	Number of Respondents	Respondents in Percentage
20-30	20	67%
31-40	9	30%
41-50	1	3%
	30	100%

Table 1 Respondents' Age

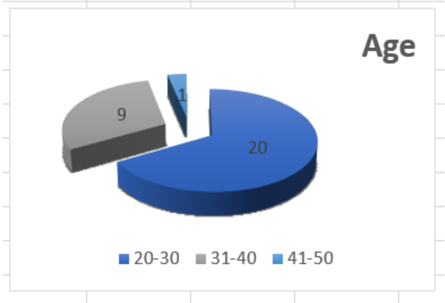
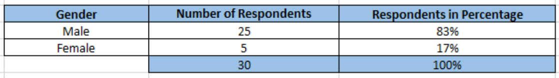


Figure 1 Respondents' Age

Table 1 and Figure 1 shows the tabulated and chart of the respondents in terms of age from the population used by the researchers. The study is composed of 30 respondents divided into two surveys. Ages 20-30 years compose of 20 respondents which has the large respondents from the population, 31-40 years of age are 9 which the 2nd biggest number of respondents, and 41-50 years of age are 1 which is the lowest number of respondents from the research.

Respondents in terms of Gender



Gender	Number of Respondents	Respondents in Percentage
Male	25	83%
Female	5	17%
	30	100%

Table 2 Gender of Respondents

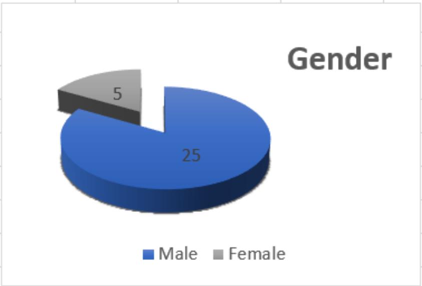
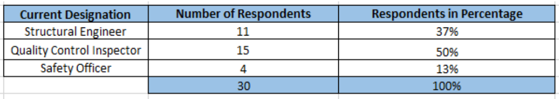


Figure 2 Gender of Respondents

Table 2 and Figure 2 Shows that there are 25 respondents for male covering 83% of population and 5 respondents for female having 8% based on the scope of works tackled in the research study.

Respondents in terms of Current Designation



Current Designation	Number of Respondents	Respondents in Percentage
Structural Engineer	11	37%
Quality Control Inspector	15	50%
Safety Officer	4	13%
	30	100%

Table 3 Respondents' Designation

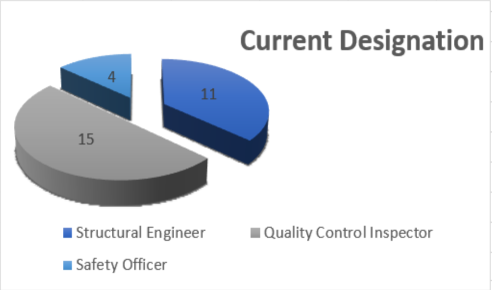
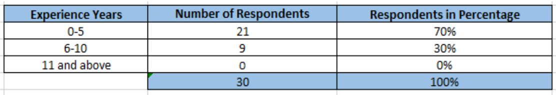


Table 3 Respondents' Designation

To respond to the researcher's questions respondents must answer the questions based on their current work designation as shown in Table 1.3 and Figure 1.3. Designation of respondents are categorized as structural engineer with 11 respondents, quality control inspector of 15 respondents, safety officer with 4 respondents.

Respondents in terms of Years of Experience



Experience Years	Number of Respondents	Respondents in Percentage
0-5	21	70%
6-10	9	30%
11 and above	0	0%
	30	100%

Table 4 Respondents' Years of Experience

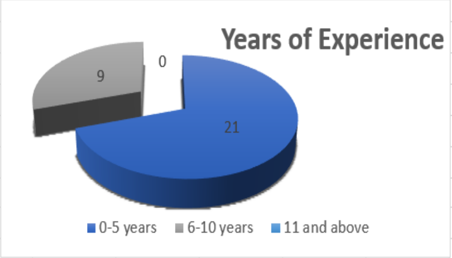
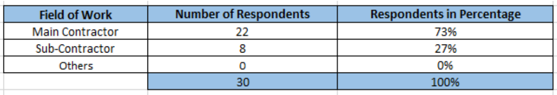


Figure 4 Respondents' Years of Experience

The respondent years of experience makes the data reliable to use for the study as the answers of the respondent are based on their years of experience working on their field of work and designation. Table 4 and Figure 4 shows that 21 have worked for 0-5 years, 9 worked for 6-10 years, and 0 who works for 11 years and above.

Respondents in terms of Type of Organization



Field of Work	Number of Respondents	Respondents in Percentage
Main Contractor	22	73%
Sub-Contractor	8	27%
Others	0	0%
	30	100%

Table 5 Respondents Type of Organization

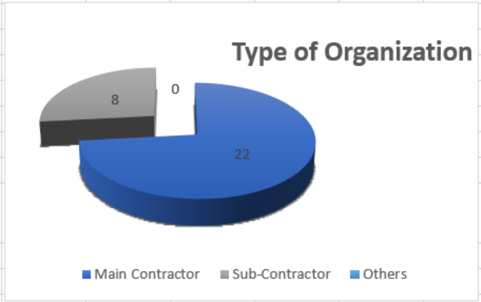


Figure 5 Respondents Type of Organization

There is different type of organization the researchers considered. In table 5 and figure 5, 22 are main contractor, and 8 is sub-contractor. Having these 2 types of categories will produce mix answers from the survey respondents.

V. CONCLUSION

As the conclusion to this research, the researchers were able to attain the goal of showing and discussing the level of significance of implementation of risk-based approach and strategies for quality control inspection in structural works, specifically on high-rise projects located in Makati City, Metro Manila. As the researchers accomplished the objectives of the study, the researchers came up with assessment and instructions acquired from the conducted survey and interview to be able to achieve, identify, and review quality assurance inspections in structural work activities. Furthermore, the researchers have also determined the frequency and severity of risks during inspections in structural works and have come up to guidelines and risk-based approach as to how to cope up with the unsafe incidents, worse, fatality, caused by negligence, lack of communication, mishandling of manpower, tools, and equipment during the inspection on structural works. Moreover, it has been proven that the results based on researchers' survey imposes different specifics which can be prioritized more to be able to conduct and strengthen risk instructions and policies, which can be empowered by safety officers, to attain a safe and strong structure. It proves that hazards of working in construction could not be prevented hundred percent and are continuously linked to project's safety management, but bringing out safety prioritization on the most critical part of construction work scope; structural works, should be then prioritize, considering the structural stage is the platform of all other field such as masonry, finishing, etc. This only partake the fact that structural work inspections can handle both quality and safety only if given the proper assessment and implementation of company guidelines, internal quality resolutions among field engineers and inspectors, and appropriate sequence of work events from pre-structural, during, and post-structural curing period of the poured structure.

VI. RECOMMENDATION

Based on the results of the survey that provide sustainable data and analysis of the answers of questionnaires, as well as the statements given by the credible respondents on the interview, this research paper makes the following recommendation to the next graduating students who wants to focus on the structural works, to further specific or breakdown the parts of structural cope. By doing so, the scope of the research will be more specific, and the number of target respondents would be narrowed down. To further improve the study, the researchers recommend narrowing down the parameters of the study. Structural works is a good scope to begin with, but entails different trade such as Rebar, Formworks, and Concreting, wherein the future researchers can possibly choose one among the mentioned structural works. In addition, it would be a stronger ground of study if researchers will involve situations correlated in an unsafe incident happened in active floors which resulted to conflict occurred during quality control inspection. This way, risk-based approach will be more centralized and specific situations shall be then analyzed together with an approach such as comparative analysis.

VII. REFERENCES

Fang, Y. (2015). Discussion on Key Points of Project Quality Control of High-rise Building. World Construction, vol. 4(4), 46-48.

Juran, J. M. (1988). Juran's Quality Control Handbook, 4th edn. McGraw-Hill. New York, NY. Makati Buildings. (n.d.). Emporis.

JB Menzies Hazards, Risk and, Structural Safety, Institution of Structural Engineers https://www.istructe.org/webtest/files/29/29b77890-c03e-48b7-a15a-b65f39137fc5.pdf

Peter G. Furst (2015). Construction Quality Management, IRMI https://www.irmi.com/articles/expert-commentary/construction-quality-management

Polat, G., Turkoglu, H., and Gurgun A. P. (2017), Identification of Material-Related Risks in Green Buildings, Procedia Engineering, vol. 196, pp. 956 – 963

Polat, G., Turkoglu, H., and Gurgun A. P.Salvi, S. S. & Kerkar, S. S. (2020). Quality Assurance and Quality Control for Project Effectiveness in Construction and Management, International Journal of Engineering and Research & Technology, vol. 9(2), 26-28. (2017), Identification of Material-Related Risks in Green Buildings, Procedia Engineering, vol. 196, pp. 956 – 963

Manoharan M. (2017). Project Management and its Effects of Quality Control in Construction Sector, 7(2)

National Building Code of the Philippines (2015). Department of Public Works, Transportation, and Communications, Philippines

Ying, C. (2010). Quality Control of Construction Projects (Thesis), University of Savonia, Varkaus, Fi.

Oğuzhan Yavuz Bayraktar (2020). World Journal of Advanced Research and Reviews: Risk management in Construction Sector

Jin, X., Zhang, G., Liu, J., Feng, Y., and Zuo, J. (2017). Major Participants in the Construction Industry and Their Approaches to Risks: a Theoretical Framework, Procedia Engineering, vol. 182, pp. 314 – 320.

Abukharmh, R. (2007). The Implementation of a Probabilistic Risk-based Inspection Approach, Faculty of Engineering and Applied Science Memorial University of Newfoundland

Soares, W. , de Vasconcelos, V., and Rabello, E. (2015). Risk-based Inspection in the Context of Nuclear Power Plant

Deqiang Liao (2020) Research on the Quality Control of Civil Engineering of High-Rise Buildings

Liang, B., Yang, X., and Liu, M. (2016) Construction Technology and Quality Control of High-rise Building, School of Sichuan, Southwest Petroleum University, China

Mr. A.K.S Priyadharsan and M.Naveen Raja (2020). International Research Journal of Engineering and Technology, Volume: 07 Issue: 05

David Arditi and H Murat Gunaydin (1997). International Journal of Project Management Vol. 15, No. 4, pp. 235-243

Kallen, MJ. (2002) Risk Based Inspection in the Process and Refining Industry, Faculty of Information Technology and Systems Technical University of Delft Delft, The Netherlands

Peter O. Akadiri, Ezekiel A. Chinyio, and Paul O. Olomolaiye (2012). Design of A Sustainable Building: A Conceptual Framework for Implementing Sustainability in the Building Sector65270035.html#utm_source=PRNewswire&utm_medium=Referal&utm_campaign=PaidPR

Mandouh Mohamad (2020). Risk-Based Prioritization Approach to Construction Inspections for Transportation Projects, Ph.D. Student, Dept. of Civil, Environmental, and Architectural

Engineering, Univ. of Kansas Totten, V. Y., Panacek, E., & Price, D. (1999). Survey Research Methodology: Designing the Survey Instrument. Air Medical Journal, 18 (1).

Yong Y.C., & Mustaffa N.E. (2013). Critical success factors for Malaysian construction projects: an empirical assessment. Construction Management and Economics, (2013) 9, 959- 978.

Isolated Iterating Fractal Independently Corresponds with Light and Foundational Quantum Problems

Blair D. Macdonald

Abstract**—**After nearly one hundred years of its origin, foundational quantum mechanics remains one of the greatest unexplained mysteries in physicists today. Within this time, chaos theory and its geometry, the fractal, has developed. In this paper the propagation behavior with iteration of a simple fractal, the Koch Snowflake, was described and analyzed. From an arbitrary observation point within the fractal set, the fractal propagates forward by oscillation—the focus of this study, and retrospectively—behind—by exponential growth from a point beginning. It propagates a potentially infinite exponential oscillating sinusoidal wave of discrete triangle bits sharing many characteristics of light and quantum entities. The models wave speed is potentially constant; offering insights into the perception and a direction of time where, to an observer when travelling at the frontier of propagation, time may slow to a stop. In isolation, the fractal is a superposition of component bits where position and scale present a problem of location. In reality this problem is experienced within fractal landscapes or fields where 'position' is only 'known' by the addition of information or markers. The quantum 'measurement problem', 'uncertainty principle', 'entanglement' and the quantum-classical interface are addressed; these are a problem of scale-invariance associated with isolated fractality. Dual forward and retrospective perspectives of the fractal model offer the opportunity of unification between quantum mechanics and cosmological mathematics, observations, and conjectures. Quantum and cosmological problems may be different aspects of the one— fractal—geometry.

Keywords**—** Entanglement, Measurement Problem, Observer, Unification

B. D. Macdonald is an economics teacher, Stockholm Sweden, phone: 0046739504740; email: blair.macdonald@taby.se

Unsteady Forced Convection Flow and Heat Transfer Past a Blunt Headed Semi-Circular Cylinder at Low Reynolds Numbers

Y. El khchine[a]*, M. Sriti[b]

[a] *Energy, Materials and Sustainable Development Laboratory, National High School of Arts and Crafts, Moulay Ismail University, Meknes, Morocco;*

[b] *Energy, Materials and Sustainable Development Laboratory, National High School of Arts and Crafts, Moulay Ismail University, Meknes, Morocco;*

*Corresponding author: Y. El khchine, email: y.elkhchine@umi.ac.ma

Abstract: In the present work, the forced convection heat transfer and fluid flow past an unconfined semi-circular cylinder is investigated. The two-dimensional simulation is employed for Reynolds numbers ranging $10 \leq Re \leq 200$, employing air ($Pr = 0.71$) as an operating fluid with Newtonian constant physics property. Continuity, momentum, and energy equations with appropriate boundary conditions are solved using the Computational Fluid Dynamics (CFD) solver Ansys Fluent. Various parameters flow such as lift, drag, pressure, skin friction coefficients, Nusselt number, Strouhal number, and vortex strength are calculated. The transition from steady to time-periodic flow occurs between Re=60 and 80. The effect of the Reynolds number on heat transfer is discussed. Finally, a developed correlation of Nusselt and Strouhal numbers is presented.

Keywords: Forced convection, semi-circular cylinder, Nusselt number, Prandtl number

References

[1] D. Chatterjee, B. Mondal, Effect of thermal buoyancy on fluid flow and heat transfer across a semicircular cylinder in cross-flow at low Reynolds numbers, Numer. Heat Transfer Part A, 67, 436–453 (2015).

[2] D. Chatterjee, B. Mondal and P. Halder, Unsteady forced convection heat transfer over a semicircular cylinder at low Reynolds numbers, Numer. Heat Transfer Part A, 63, 411–429 (2013).

[3] D. Chatterjee, B. Mondal, Mixed Convection Heat Transfer From an Equilateral Triangular Cylinder in Cross Flow at Low Reynolds Numbers, Heat Transfer Engineering, 36(1), 123-133 (2015).

[4] A. K. Sahu, R. P. Chhabra, and V. Eswaran, "Effects of reynolds and prandtl numbers on heat transfer from a square cylinder in the unsteady flow regime," Int. J. Heat Mass Transfer, vol. 52, pp. 839–850 (2009).

[5] A. Sharma and V. Eswaran, "Heat and fluid flow across a square cylinder in the two-dimensional laminar flow regime," Numer. Heat Transfer Part A Appl. 45, 247 (2004).

[6] A. K. Dhiman, R. P. Chhabra, and V. Eswaran, "Flow and heat transfer across a confined square cylinder in the steady flow regime: Effect of peclet number," Int. J. Heat Mass Transfer, vol. 48, pp. 4598–4614, 2005.

[7] A. Chandra and R. P. Chhabra, "Influence of power-law index on transitional Reynolds numbers for flow over a semi-circular cylinder," Appl. Math. Model. 35, 5766 (2011).

[8] S. A. Patel and R. P. Chhabra, "Steady flow of Bingham plastic fluids past an elliptical cylinder," J. Non-Newtonian Fluid Mech. 202, 32 (2013).

[9] S. Bhowmick, Md. M. Molla, and L.S. Yao, "Non-Newtonian mixed convection flow along an isothermal horizontal circular cylinder", Numerical Heat Transfer, Part A, 66: 509–529, (2014)

[10] A. Chandra and R. P. Chhabra, "Flow over and forced convection heat transfer in Newtonian fluids from a semicircular cylinder," Int. J. Heat Mass Transfer, vol. 54, pp. 225–241, 2011.

[11] A.P. Pawar, S. Sarkar, S.K. Saha, Forced convective flow and heat transfer past a blunt headed cylinder with corner modification, Phys. Fluids 33, 103106 (2021)

[12] A. Kumar and A. Dhiman, Laminar Flow and Heat Transfer Phenomena Across a Confined Semicircular Bluff Body at Low Reynolds Numbers, Heat Transfer Engineering, 36(18):1540–1551, 2015

[13] H. Kapadia, A. Dalal, S. Sarkar, Forced convective flow and heat transfer past an unconfined blunt headed cylinder, Numerical Heat Transfer, Part A: Applications, 72(5), 372-388, 2017

[14] B. N. Rajani, A. Kandasamy, and S. Majumdar, "Numerical simulation of laminar flow past a circular cylinder," Appl. Math. Model. 33, 1228 (2009).

[15] A. Pal Singh Bhinder, S. Sarkar, and A. Dalal, "Flow over and forced convection heat transfer around a semi-circular cylinder at incidence," Int. J. Heat Mass Transfer 55, 5171 (2012).

[16] S. Bhadauriya, H. Kapadia, A. Dalal, and S. Sarkar, "Effect of channel confinement on wake dynamics and forced convective heat transfer past a blunt headed cylinder," Int. J. Therm. Sci. 124, 467 (2018).

[17] W. Zhang and R. Samtaney, "Low-Re flow past an isolated cylinder with rounded corners," Comput. Fluids 136, 384 (2016).

[18] T. Ambreen and M. H. Kim, "Flow and heat transfer characteristics over a square cylinder with corner modifications," Int. J. Heat Mass Transfer 117, 50 (2018).

[19] S. Miran and C. H. Sohn, "Numerical study of the rounded corners effect on flow past a square cylinder," Int. J. Numer. Methods Heat Fluid Flow 25, 686 (2015).

[20] M. M. Alam, T. Abdelhamid, and A. Sohankar, "Effect of cylinder corner radius and attack angle on heat transfer and flow topology," Int. J. Mech. Sci. 175, 105566 (2020).

[21] P. Dey and A. K. Das, “Heat transfer enhancement around a cylinder—A CFD study of effect of corner radius and Prandtl number,” Int. J. Chem. React. Eng. 14, 587 (2016).

[22] F. Zafar and M. M. Alam, “Flow structure around and heat transfer from cylinders modified from square to circular,” Phys. Fluids 31, 083604 (2019).

LncRNA-miRNA-mRNA Networks Associated with BCR-ABL T315I Mutation in Chronic Myeloid Leukemia

Adenike Adesanya, Nonthaphat Wong, Xiang-Yun Lan, Shea Ping Yip, Chien-Ling Huang

Department of Health Technology and Informatics,
The Hong Kong Polytechnic University, Hong Kong, China

Abstract—

Background: The most challenging mutation of the oncokinase BCR-ABL protein T315I, which is commonly known as the "gatekeeper" mutation and is notorious for its strong resistance to almost all tyrosine kinase inhibitors (TKIs), especially imatinib. Therefore, this study aims to identify T315I-dependent downstream microRNA (miRNA) pathways associated with drug resistance in chronic myeloid leukemia (CML) for prognostic and therapeutic purposes.

Methods: T315I-carrying K562 cell clones (K562-T315I) were generated by CRISPR-Cas9 system. Imatinib-treated K562-T315I cells were subjected to small RNA library preparation and next generation sequencing. Putative lncRNA-miRNA-mRNA networks were analyzed with (i) DESeq2 to extract differentially expressed miRNAs, using P_{adj} value of 0.05 as cut-off, (ii) STarMir to obtain potential miRNA response element (MRE) binding sites of selected miRNAs on lncRNA H19, (iii) miRDB, miRTarbase, and TargetScan to predict mRNA targets of selected miRNAs, (iv) IntaRNA to obtain putative interactions between H19 and the predicted mRNAs, (v) Cytoscape to visualize putative networks, and (vi) several pathway analysis platforms – Enrichr, PANTHER and ShinyGO for pathway enrichment analysis. Moreover, mitochondria isolation and transcript quantification were adopted to determine the new mechanism involved in T315I-mediated resistance of CML treatment.

Results: Verification of the CRISPR-mediated mutagenesis with digital droplet PCR detected the mutation abundance of ⩾ 80%. Further validation showed viability of ⩾90% by cell viability assay, and intense phosphorylated CRKL protein band being detected with no observable change for BCR-ABL and c-ABL protein expressions by Western blot. As reported by several investigations into hematological malignancies, we determined a 7-fold increase of H19 expression in K562-T315I cells. After imatinib treatment, a 9-fold increment was observed. DESeq2 revealed 171 miRNAs were differentially expressed K562-T315I, 112 out of these miRNAs were identified to have MRE binding regions on H19, and 26 out of the 112 miRNAs were significantly downregulated. Adopting the seed-sequence analysis of these identified miRNAs, we obtained 167 mRNAs. 6 hub miRNAs (hsa-let-7b-5p, hsa-let-7e-5p, hsa-miR-125a-5p, hsa-miR-129-5p, and hsa-miR-372-3p) and 25 predicted genes were identified after constructing hub miRNA-target gene network. These targets demonstrated putative interactions with H19 lncRNA and were mostly enriched in pathways related to cell proliferation, senescence, gene silencing, and pluripotency of stem cells. Further experimental findings have also shown the up-regulation of mitochondrial transcript and lncRNA MALAT1 contributing to the lncRNA-miRNA-mRNA networks induced by BCR-ABL T315I mutation.

Conclusions: Our results have indicated that lncRNA-miRNA regulators play a crucial role not only in leukemogenesis, but also in drug resistance, considering the significant dysregulation and interactions in K562-T315I cell model generated by CRISPR-Cas9. *In silico* analysis has further showed that lncRNAs H19 and MALAT1 bear several miRNA complementary sites. This implies that they could serve as a sponge, hence sequestering the activity of the target miRNAs.

Keywords: chronic myeloid leukemia, imatinib resistance, lncRNA-miRNA-mRNA, T315I mutation

Dual Biometrics Fusion Based Recognition System

[1]Prakash, [2]Vikash Kumar, [3]Vinay Bansal, [4]LN Das

[1]Final Year, Bachelor of Technology, Delhi Technological University
[2] Final Year, Bachelor of Technology, Delhi Technological University
[3] Final Year, Bachelor of Technology, Delhi Technological University [4]Professor, Delhi Technological University
Email: [1] prakashahirwar0402@gmail.com, [2]Vvikash157@gmail.com, [3]Vinay.bansal@gmail.com, [4]lndas@dce.ac.in
Contact: [1](+91) 9289842217, [2](+91) 9667586905, [3](+91) 8527944311, [4](+91) 88265 70638

Abstract: Dual biometrics is a subpart of multimodal biometrics, which refers to the use of a variety of modalities to identify and authenticate persons rather than just one. We limit the risks of mistakes by mixing several modals, and hackers have a tiny possibility of collecting information. Our goal is to collect the precise characteristics of iris and palmprint, produce a fusion of both methodologies, and ensure that authentication is only successful when the biometrics match a particular user. After combining different modalities, we created an effective strategy with a mean DI and EER of 2.41 and 5.21, respectively. A biometric system has been proposed.

Index terms: Multimodal, fusion, palmprint, Iris,EER,DI

1. INTRODUCTION

Biometrics with a single mode recognition faces several issues such as data noise, results fluctuation, limited degrees of freedom, and high mistake rates because single biometrics are used,. The purpose of multiple biometric identification systems is to solve these issues.We believe that the biometric system is preferable to traditional ways since it is more precise, personalised, and harder to produce. We are aware that using this method limits performance. When it comes to biometrics, we have to be extremely careful. For this model, we selected two of the most popular but most secure and recognisable modalities: palmprint and iris. Palmprint identification was done using a minutiae-based feature extraction methodology, while iris identification was done using an old but effective method called Daugman Algorithm fusion.Both features are retrieved from the modalities first, and then a single feature is formed (or both features are blended into one) before making the final judgement.

We claim that when compared to alternative fusion approaches such as score level fusion or decision level fusion, the primary drawback is that information loss is substantially higher due to the translation of one-dimensional data into a single match score. Our feature level fusion, on the other hand, combines characteristics that contain more decisive and unique information. Cancellable Biometrics solution is more secure and protects modalities against assaults such as dictionary attacks and brute force attacks.
It's a method of preserving the user's originality by creating phoney biometric IDs in one or both directions. Using a user-specific key data set or a transformation function.
The canonical correlation analysis is used to combine these characteristics (CCA). We separate the vectors into two categories. The correlating criterion function and its canonical correlation characteristics are then introduced..

2. Related Work

For this particular part , we would perform a grasp of how things function in order to point out the differences in process More than one biometric approach has been thoroughly examined in order to provide reliable results..

Score level fusion, we may say, is an excellent choice for multi-modal biometrics since it is simple to integrate match scores. In that vein, we've suggested quality score fusion models with weights and confidence factors. Recently, researchers used the Back-track Search Optimization Algorithm to fuse the iris and palmprint..

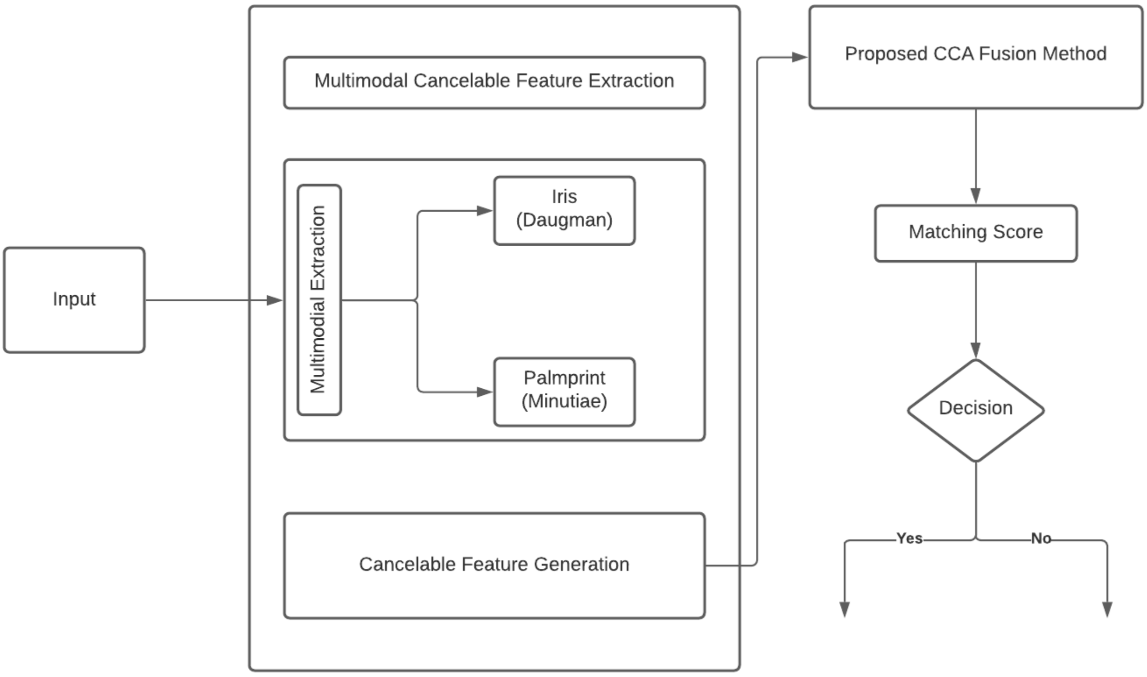
Despite the fact that score level fusion is far more difficult than score fusion, it is more theoretically investigated due to its lower flexibility than match scores. The inconsistency in type and size of separate feature sets causes problems with feature level fusion..

In fusions for eye and palm modalities were recently done without using a single feature vector..

Individual features are extracted using local texture descriptors such as LBP, and a single feature template is generated as a result. In order to build a homogenous biometric vector, a log-Gabor based filter was utilised to encode palm and eye characteristics. The bulk of multimodal biometric fusion approaches employ linear fusion methods, which are ineffective for fusing data from many modalities..

Aside from the performance, the biometric system's privacy is protected. The majority of extant biometrics systems employ one of two methods: bio-hashing or non-invertible transformation. The bio-hashing approach is similar to password-salting in that it modifies the primary biometrics templates using a pin or a password..

Fig. (1) Diagram of the our biometric approach.


Multimodal Cancelable Feature Extraction
Input
Multimodial Extraction
Iris
(Daugman)
Palmprint
(Minutiae)
Cancelable Feature Generation
Proposed CCA Fusion Method
Matching Score
Decision
Yes
No

3. PROPOSED METHODOLOGY

We have proposed a multi-modal biometrics system that generates crucial pictures-based characteristics and combines them with feature fusion utilising Canonical correlation analysis. The technology is used to do recognition based on biometric traits such as palm and eyes.

The CCA fusions strategy is an useful method for extracting information from numerous biometrics while also ensuring the security of our system. We select the eyes and palms as authentication methods since they are the most prevalent and may be utilised for biometric authentication. Figure 1 depicts the architecture of the same framework.

.

A. Feature Extraction

For eye extraction, we employ a multi-modal feature, and for palm-print extraction, we use two feature-extraction approaches, primarily the daugman algorithm and minutiae-based extraction.In this procedure, specific modalities are extracted.

(a) Keys are analysed in the iris and palmprint modalities, then sent through a modified CCA feature fusion model to yield multimodal features.

Eyes, a crucial biometrics strategy that relies on high-level recognition. Iris is a one-of-a-kind piece of information. The Daug-man algorithms are used by several eye recognition systems. Using the Integro Differential Daugman Operator, locate the iris and eyelids..

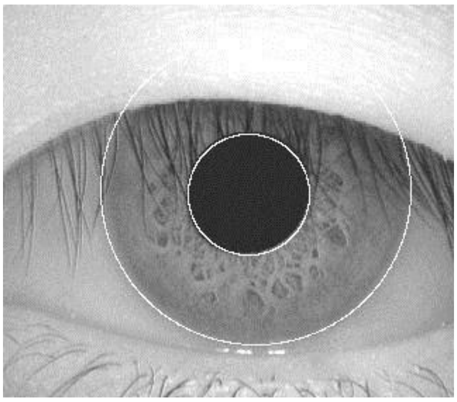
For finding the iris and pupil areas, as well as the eyelids, Daugman employs an intregro differencial operator. The operator is defined as

$$Max_{(r,x0,y0)} \left| G_\sigma(r) * \frac{\partial}{\partial r} \oint I(\frac{(x,y)}{2\pi r}) ds \right| \quad (1)$$

where I(x, y) is the ocular image, r is the search radius, G(r) is the Gaussian smoothing function, and s is the circle shape defined by r, x0, y0. This work converts a cartesian eye picture into a size-invariant, normalised, and non concentric polar system Ii(x,y) Ii(r,), where x and y are calculated as follows:

$$x = (1 - r)(xp(\theta)) + rxi(\theta)$$
$$y = (1 - r)(yp(\theta)) + ryi(\theta)$$

The search from the pupil is strated by the algorithm to determine the greatest changing pixel value, where I(x,y) represents the eye picture. (particular derivative)



Fig(2).eye showing segment

The Hough-Transform is a fundamental system visioning technique for determining the parameters of a given object..

The radii and centres coordinates of the ocular area are calculated using the circular Hough transform..

$$(x -_{i} x)^2 +_{i} (y - y)^2 + r^2$$

(2)

The hamming-distance may be used to determine how similar two bits are. We may see if bit-patterns are from the same or distinct iris or eyes by using hamming -distance.

The Hamming Distance may be defined as the total of two disagreeing bits when comparing bit patterns like X and Y.

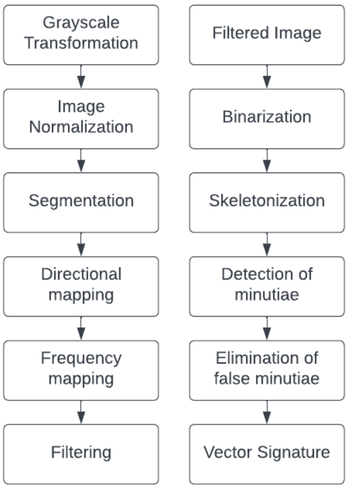
$$HD = \frac{1}{N}\sum_{j=1}^{N} X_j (XOR) Y_j$$

(3)

The misalignments in the normalised iris pattern produced by rotational variations during imaging were rectified using DAUGMAN. As a result, we get Ii Qi, Khi for the normalised iris area..

The pre-processing of the palm-print picture to increase its quality and the extraction of its signature are two crucial aspects of the palmprint identification algorithm.

One of the most crucial phases of the algorithm is pre-processing. We claim it allows us to improve the image in order to make the work easier in the next stage and to optimise the image processing.



Fig(3). Different pre-processing phases

Figure 1 depicts the various pre-processing stages (3).

The technique shown in the following picture was used to extract biometric data (the biometric data about the palmprint are the minutiae). Normalization is utilised for standard intensity values in the provided picture.

- Averaging

$$M = \frac{1}{n*m}\sum_{i=0}^{n-1} \sum_{j=0}^{m-1} I(i,j)$$

(4)

I I j) is the pixel value I j), M is the image's average value, and m, n are the image's dimensions..

- Variance Calculation

$$V = \frac{1}{n*m}\sum_{i=0}^{n-1} \sum_{j=0}^{m-1} (I(i,j) - M)^2$$

(5)

V is the variance of the image.

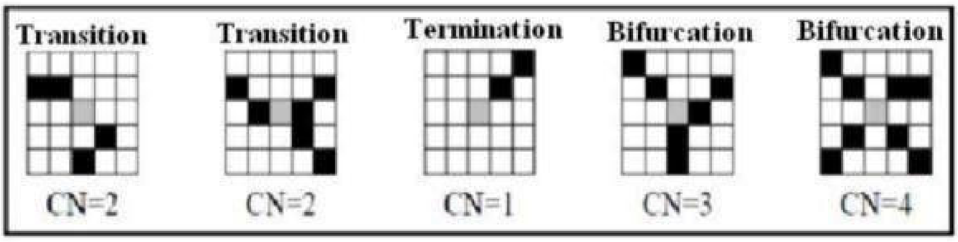
For eliminating the image's edge, we employ segmentation. We do this by calculating the grey area variance. And we divide our image into two sub-blocks (W w) of different sizes, with the variance determined for each block using the formula (2) .

The method we're employing is the Crossing-Number method (CN). It's a straightforward procedure. Using a connection of 8 neighbours (window 3 3) the minutiae are inspected for the pixels in the picture of the palmprint.

P1	P2	P3
P8	P	P4
P7	P6	P5

Thus, for a pixel P belonging to a streak (that is to say of value 1), the CN can take five values (figure 6):

- CN(P) = 0 : It is an isolated pixel, we do not take into account it because even if this type of minutia exists, it is very rare and in the general case it is due to a noise residue.
- CN(P) = 1 : It is a candidate for a termination
- CN(P) = 2 : This is the most common case, it is a pixel that is on a streak, there are no minutiae in this case
- CN(P) = 3 : A triple bifurcation candidate
- CN(P) = 4 : A quadruple bifurcation, this type is quite rare and it is probably due to noise.



$CN(P) = \frac{1}{2}\sum_{i=1}^{8} \quad |P_i - P_{i-1}|$ where P8=P0 $P_i \in \{0, 1\}$

We can accomplish feature fusion between two modalities by comparing the two Signatures, which is stored in the Signature vector. Now we'll go on to the next stage, which is feature unification.

B. Feature Unification

We produce a single feature vector using this feature fusion approach, which is a combination of the two features or feature vectors provided by the user. This method is more efficient and discriminatory than previous methods. CCA FUSE uses a Canonical Correlation Analysis (CCA) based approach for feature level fusion. Then it receives training and test data from the modalities X and Y.

$$S = \begin{pmatrix} \frac{cov(x)}{cov(y,x)} & \frac{cov(x,y)}{cov(y)} \end{pmatrix} = \begin{pmatrix} S_{xx}S_{xy} \\ S_{yx}S \end{pmatrix} \qquad (6)$$

and combine these into a single feature package. CCA generates a single feature based on the interdependence of the dual vector used as input (Canonical Correlation Analysis).

This interdependence between two vectors was created using statistical approaches .

3.2 Combination of Features CCA is a non-profit organisation that aims to (Canonical Correlation Analysis) We get a single feature vector from the user's input, which is a combination of the two features or feature vectors. This approach is much more efficient and discriminating than previous methods.

The most often used approach for analysing the relationships between two pairs of variables is CCA (Canonical Correlation Analysis).

. $X \in R^{p*n}$ and $Y \in R^{q*n}$ 275 These are a pair or set of two matrices that each have n number of training features and feature vectors derived from two separate modalities. Let's imagine there are n samples for each of the (p+q) traits that are being considered.

However, determining the relationship or connection between these two sets of feature vectors from this matrix is difficult due to the fact that the correlation between both sets of characteristics may not follow a consistent pattern (Krzanowski, 1988) .

Canonical Correlation Analysis aims to find the linear combinations Y*=WTy Y and X*=WTX X that improve the pair-wise linkages between the two data sets:

$corr(X^*, Y^*) = cov(X^*, Y^*)$, where

$corr(X^*, Y^*) = W^T S_{xy} W_y$,
$var(X^*) = W^T{}_x S_{xx} W_x$ and
$var(Y^*) = W^T{}_y S_{yy} W_y$,

Augmentation is accomplished using Lagrange multipliers, which maximise the covariance between Y* and X* under the constraints varX*=varY*285=1. Then we solve the eigen value equations, and the matrices that result are referred to as Transformation Matrices Wx and Wy, respectively.

$$S^{-1}{}_{xx}\, S_{xy}\, S^{-1}{}_{yy}\ \ S_{yx}\ \ W_x = \wedge\ 2W_x$$

$$S^{-1}{}_{yy}\, S_{yx}\, S^{-1}{}_{xx}\ \ S_{xy}\ \ W_y = \wedge\ 2W_y \qquad (7)$$

The eigenvectors are 'Wy' and 'Wx,' and the squares of the canonical correlations or diagonal matrix of eigenvalues is 2.

$$d = rank(S_{xy}) \leq min(n, p, q) \qquad (8)$$

This equation is used to extract the number of non-zero eigen values. Sxx Rp*p and Syy Rq*q are the covariance (within-set) matrices of X and Y, respectively, while Sxy Rp*q is the covariance (between-sets) matrix (accounting for Syx=STxy). S, the all-inclusive matrix (p+q)*(p+q) covariance matrix, stores or accommodates all information on interrelationships or linkages.

organized in diminishing direction, λ1 ≥ λ1 ≥ ...

≥ λd. ‘W_y’ and ‘W_x’ are transformation matrices including sorted or ordered eigen vectors akin to or equivalent to nonzero eigen values? The canonical variates Y, X R nd 290 have been recognised.

The Transformed Data, which is essentially a sample of the covariance matrix in Eq. (6), will take the form given below.

:

$$S^* = \left[\begin{array}{cccccccccc} 1 & 0 & \dots & 0 & | & \lambda_1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & | & 0 & \lambda_2 & \dots & 0 \\ \vdots & & \ddots & & | & \vdots & & \ddots & 0 \\ 0 & 0 & \dots & 1 & | & 0 & 0 & \dots & \lambda_d \\ - & - & - & - & - & - & - & - & - \\ \lambda_1 & 0 & \dots & 0 & | & 1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 & | & 0 & 1 & \dots & 0 \\ \vdots & & \ddots & & | & \vdots & & \ddots & 0 \\ 0 & 0 & \dots & \lambda_d & | & 0 & 0 & \dots & 1 \end{array}\right]$$

The variates (canonical) exhibit nonzero connectivity or correlation solitary on their conforming indices, as seen in the matrix above.

The matrices with zeros elsewhere and ones on the main diagonal in the bottom right and upper left corners show that the variates (canonical) are not interrelated or associated inside each data-set. According to (Sun et al., 2005), feature fusion or fusion at the feature level is also done by summing or concatenation of the updated feature vector..

$$Z_1 = \left(\frac{X^*}{Y*}\right) = \left(\frac{W^T{}_x X}{W^T\ \ {}_y Y}\right) \qquad (9)$$

$$Z_2 = X^* + Y^* = W^T{}_x X = \left(\frac{W_x}{W_Y}\right)^T \left(\frac{X}{Y}\right) \quad (10)$$

The C.C.D.Fs are Z2 and Z1, respectively (Canonical Correlation Discriminant Features).

4. EXPERIMENTAL RESULTS AND DISCUSSIONS

Experimentation on yardstick or benchmark datasets is used to evaluate the expected or suggested multimodal biometric system, and the outcomes/results are compared to state-of-the-art approaches. Experimental Validation and its specifics are described further down.

The following are the specifics of the experiment validation: CASIA Iris-V1 (CASIA Iris Image Database Version 1.0) and MMU2 [38]. A.Experimental Design and Database CASIA Iris-V1 (CASIA Iris Image Database Version 1.0) and MMU2 [38]. These are the two benchmarks or datasets on which Experiment is run. This is a collection of two iris datasets..

- CASIA is the acronym for CASIA Iris Image Database Version 1.0 (CASIA Iris V1 i.e. CASIA Iris Image Database Version 1.0). This dataset contains 756 iris pictures acquired from 108 eyes. Images are then saved or stored in bitmap format, i.e. BMP, with a resolution of 320*280 pixels. Then our Database DB – 1 is assembled or produced manually after the selection of 1000 photographs from 200 individuals, each of which contains 5 images of the left side of the eye. Currently, three photos are used in the model's training, with the other two stored elsewhere. For the model, two photographs are utilised or are being used as Test Images. These three photographs were chosen at random as training images and the other two as test images.

Database of palmprints Palmprint is utilised as our modality or palmprint datasets, and one of our datasets was used for Palmprint.,

CASIA-PalmprintV5 (or CASIA Palmprint-Image with Database-Version 5) is a collection of two thousand palmprint pictures from fifty different participants. Volunteers supplied five photographs of each palm, resulting in a total of 40 images of volunteer palmprints. A total of eight palms . The images of all the palmprints in the collection are BMP files with an 8-bit grey level.

The photos have a resolution of 328*356 pixels. Currently, three photos are used in the model's training, with the other two stored elsewhere. Two photographs have been used or are being used as Test Images.

Biometrix has released another database, DB-2, which has 600 palmprint pictures. There are 120 palms in all, with five photos of each palm. All of the palmprint images in the collection are in BMP format with an 8-bit grey level. The photos have a resolution of 328*356 pixels. Currently, three photos are used in the model's training, with the other two stored elsewhere. Two photographs have been used or are being used as Test Images.

The training and testing were done with MATLAB 2016 and an NVIDIA GTX 1650 with 12GB RAM. To get our final result, we average out the results that we show in the output.

B. Performance Metrics

The first is EER, which stands for equal error rate, and the second is DI, which stands for decidibility index, which is used to measure the effectiveness of our model.

To measure the effectiveness of our biometric system, we are comparing two of the most efficient fusion approaches, which we have taken utilising two rates, FMR (false match rate) and FNMR (false non-match rate), as well as the FAR (false acceptance rate).

We have taken into consideration EER, which stands for Equal Error Rate, as previously stated. .It's the error rate at which our calculated FAR crosses FRR, implying that it's equal to that. Also, as previously indicated, the decidability index (DI) plays a significant part in determining the efficiency of our approach.

DI is used to determine the separation or distance between the original and duplicate distributions. DI is calculated as follows:

$$DI = \frac{\mu_g - \mu_i}{\sqrt{(\sigma^2{}_g + \sigma^2{}_g)/2}} \qquad (11)$$

The mean and standard deviation of the original and duplicate score distributions are g and im, respectively. We may also find the IP, which stands for identity performance, by utilizing the RI, which stands for Recognition index. The highest values of each topic are used to calculate it. Furthermore, two curves, CMC and DET, have been drawn to highlight the relationship between FRR and FAR.

C. Quantitative Analysis

We have taken different extraction methods to compare with our proposed one and compare the DI and EER respectively, and we have also compared our fusion methods with old traditional methods. After finding performance metrics, we have to proceed to quantitative analysis of our multimodal system so that we can easily find how well our model is performing. Below is a detailed study and comparison..

1) Comparison of Different Feature Extraction Methods

To compare the data and assess our efficiency, we used LBP and Gabor, which are two older and less efficient approaches that give us an indication of the efficiency of numerous feature extraction methods.

As previously said, we have taken two modalities, namely iris and palmprint, and created a table with them. the results which can be found below,We can see that our proposed fWe have taken two distinct databases for both modalities and done feature fusion on them for better understanding and comparison. We have also taken two different databases for both modalities and performed feature fusion on them for better understanding and comparison.

2) Comparison of Fusion Methods

We must now compare fusion strategies to other approaches. We are comparing the proposed fusion method to traditional GABOR and LBP fusion methods, in which individual features are extracted first using iris and palmprint based feature extraction methods, and then fusion is performed using CCA based fusion methods. The results show that CCA based fusion outperforms traditional methods by a significant margin.

PRIVACY ANALYSIS

Our fusion approach has concentrated on many such factors as non-invertibility, repeatability, unlikability, and revocability, guaranteeing that user data is safe and secure.

A. Non Invertibility

It protects biometric data against an assault on the template, and in the worst-case scenario, both the template and the matrix are available to the attacker.

B. Revocability

It guarantees that templates aren't connected to one another or have been used by the same user before. To distinguish the new templates from the old ones, we use separate keys for both the original and new templates.

Table 1. Individual biometric comparison based on different Fusion Algorithms

TABLE (I)

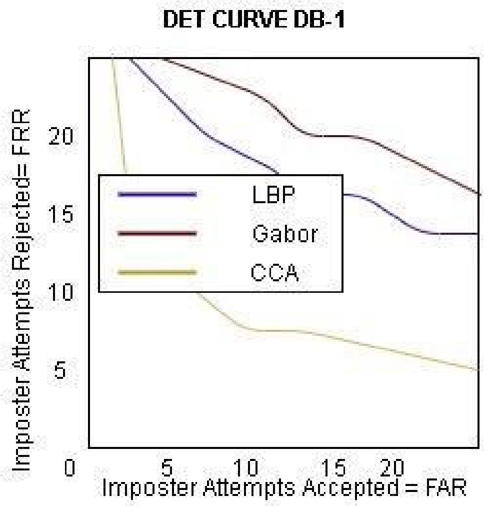
Various feature extraction algorithms comparison for Palmprint

	DB-1		DB-2	
architecture	DI	EER	DI	EER
LBP	1.69 ± 0.07	22.29 ± 0.32	1.22 ± 0.05	21.79 ± 0.43
Gabor	1.6 ± 0.32	20.32 ± 0.50	1.27 ± 0.36	23.89 ± 0.29
proposed Daugman	5.36 ± 0.31	3.94 ± 0.60	4.72 ± 0.28	3.27 ± 0.37

TABLE (II)

Various feature extraction algorithms comparison for Iris

	DB-1		DB-2	
architecture	DI	EER	DI	EER
LBP	3.11 ± 0.68	14.13 ± 0.21	2.55 ± 0.05	19.31 ± 0.32
Gabor	3.21 ± 0.11	9.21 ± 0.28	3.42 ± 0.09	24.38 ± 0.40
proposed Daugman	7.28 ± 0.41	4.95 ± 0.37	6.65 ± 0.43	4.16 ± 0.51



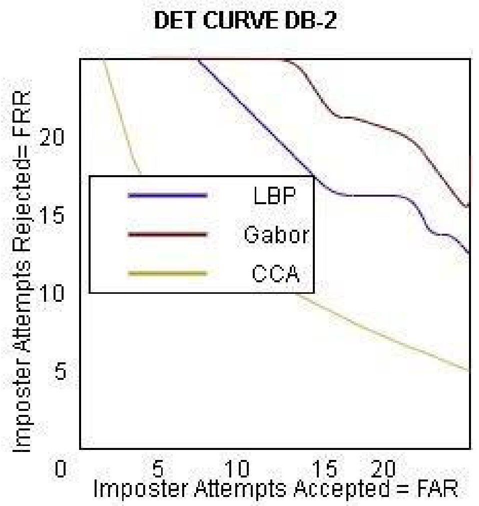


Table 3. Results for various Fusion Methods

	DB-1		DB-2	
Architecture	DI	EER	DI	EER
Gabor	4.12 ± 0.20	13.2 ± 0.49	2.91 ± 0.26	16.1 ± 1.28
LBP	3.22 ± 0.34	12.96 ± 0.56	2.32 ± 0.19	13.1 ± 0.71
proposed CCA	2.47 ± 0.83	6.05 ± 0.02	2.23 ± 0.69	5.19 ± 0.08

SECURITY ANALYSIS

A . Unlinkability

It assures that all templates created are unlinkable, even if they use different keys, adding a layer of protection to templates used in various databases.

B. Attacks via Record Multiplicity

We are preventing arm assaults in the proposed fusion by reshuffling intensity values across the neighborhood to prevent these attacks, which means the hacker obtains knowledge about different templated with their parameters to construct the original template.

C . Brute Force Attack

It implies that the hacker is ignorant of the biometric template that we use, so they try a variety of combinations to match the template, but we use such high accuracy in our process that these kind of assaults are extremely difficult to carry out.

D. Dictionary Attack

The hacker has some understanding of the template and while they may not have tried every potential combination, they are familiar with the transformation method.

,Now, a significantly smaller number is required. They attack the database using a threshold-based approach.

CONCLUSION

We're going to use two separate modalities in this strategy and then fusing them together. We're utilising two of the most frequent modalities, iris and palmprint, and we're using a daugman and minutae-based extraction approach to locate relevant features so that we can execute a feature fusion on both of them. We're fusing using a CCA-based fusion approach, and we've calculated EER and DI for both modalities and our fusion method. For better results, we utilised two databases. We've also considered the privacy and security implications of our project, such as brute force assaults and dictionary attacks, among other things. Our EER for fusion is 6.05, and our DI is 2.4, which is significantly better than previous fusion approaches.

REFERENCES

[1] A. K. Jain and K. Nandakumar, “Biometric authentication: System security and user privacy,*Computer*, vol. 45, no. 11, pp. 87–92, Nov. 2012.

[2] N. K. Ratha, J. H. Connell, and R. M. Bolle, “Enhancing security and privacy in biometrics-based authentication systems,” *IBM Syst. J.*, vol. 40, no. 3, pp. 614–634, Apr. 2001.

[3] K. Nandakumar and A. K. Jain, “Biometric template protection: Bridg ing the performance gap between theory and practice,” *IEEE Signal Process. Mag.*, vol. 32, no. 5, pp. 88–100, Sep. 2015.

[4] V. M. Patel, N. K. Ratha, and R. Chellappa, “Cancelable biometrics: A review,” *IEEE Signal Process. Mag.*, vol. 32, no. 5, pp. 54–65, Sep. 2015.

[5] J. K. Pillai, V. M. Patel, R. Chellappa, and N. K. Ratha, “Secure and robust iris recognition using random projections and sparse rep resentations,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 9, pp. 1877–1893, Sep. 2011.

[6] M. Gomez-Barrero, C. Rathgeb, G. Li, R.

Ramachandra, J. Galbally, and C. Busch, "Multi-biometric template protection based on bloom filters," *Inf. Fusion*, vol. 42, pp. 37–50, Jul.2018.

[7] A. T. B. Jin, D. N. C. Ling, and A. Goh, "BioHashing: Two factor authentication featuring palmprint data and tokenised ran dom number," *Pattern Recognit.*, vol. 37, no. 11, pp. 2245–2255, Apr. 2004.

[8] Y.-L. Lai *et al.*, "Cancellable iris template generation based on indexing-first-one hashing," *Pattern Recognit.*, vol. 64, pp. 105–117, Apr. 2017.

[9] Z. Jin, J. Y. Hwang, Y.-L. Lai, S. Kim, and A. B. J. Teoh, "Ranking based locality sensitive hashing-enabled cancelable biometrics: Index of-max hashing," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 2, pp. 393–407, Feb. 2018.

[10] M. Sultana, P. P. Paul, and M. L. Gavrilova, "Social behavioral informa tion fusion in multimodal biometrics," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 48, no. 12, pp. 2176–2187, Dec. 2018.

[11] A. M. P. Canuto, F. Pintro, and J. C. Xavier-Junior, "Investigating fusion approaches in multi-biometric cancellable recognition," *Expert Syst. Appl.*, vol. 40, no. 6, pp. 1971–1980, 2013.

[12] G. S. Walia, S. Rishi, R. Asthana, A. Kumar, and A. Gupta, "Secure multimodal biometric system based on diffused graphs and optimal score fusion," *IET Biometrics*, vol. 8, no. 4, pp. 231–242, Jul. 2019.

[13] T. Chugh, K. Cao, and A. K. Jain, "Palmprint spoof buster: Use of minutiae-centered patches," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 9, pp. 2190–2202, Sep. 2018.

[14] J. Galbally, S. Marcel, and J. Fierrez, "Image quality assessment for fake biometric detection: Application to iris, palmprint, and face recognition," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 710–724

[15] J. Daugman, "The importance of being random: statistical principles of iris recognition," Pattern Recognit., vol. 36, no. 2, pp. 279–291, 2003.

[16] D. Maltoni, D. Maio, A. K. Jain, and S. Prabhakar, Handbook of Palmprint Recognition. New York: Springer Verlag, Jun. 2003.

[17] R. Sanchez-Reillo, C. Sanchez-Avila, and A. Gonzales-Marcos, "Biometric identification through hand geometry measurements," IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 10, pp. 1168–1171, Oct. 2000.

[18] Belguechi, Rima, Estelle Cherrier, Christophe Rosenberger, and Samy Ait-Aoudia. ”Operational bio-hash to preserve privacy of palmprint minutiae templates." IET biometrics 2, no. 2 (2013): 76-84.

[19] Daugman, John G. ”High confidence visual recognition of persons by a test of statistical independence." IEEE transactions on pattern analysis and machine intelligence 15, no. 11 (1993): 1148-1161.

[20] Sun, Zhenan, and Tieniu Tan. ”Ordinal measures for iris recognition." IEEE Trans. Pattern Anal. Mach. Intell. 31, no. 12 (2009): 2211-2226.

[21] De Marsico, Maria, Michele Nappi, and Daniel Riccio. ”Noisy iris recognition integrated scheme." Pattern Recognition Letters 33, no. 8 (2012): 1006-1011.Haghighat, Mohammad, Mohamed Abdel-Mottaleb, and Wadee Alhalabi. ”Discriminant correlation analysis: Real-time feature level fusion for multimodal biometric recognition." IEEE Trans. Inf. Forensics Security 11, no. 9 (2016): 1984-1996.

[22] Walia, Gurjit Singh, Shivam Rishi, Rajesh Asthana, Aarohi Kumar, and Anjana Gupta. ”Secure multimodal biometric system based on diffused graphs and optimal score fusion." IET Biometrics (2019).

[23] Quan, Feng, Su Fei, Cai Anni, and Zhao Feifei. ”Cracking cancelable palmprint template of Ratha." In 2008 International Symposium on Computer Science and Computational Technology, vol. 2, pp. 572-575. IEEE, 2008.

[24] Shin, Sang Wook, Mun-Kyu Lee, Daesung Moon, and Kiyoung Moon. ”Dictionary attack on functional transform-based cancelable palmprint templates." ETRI journal 31, no. 5 (2009): 628-630.

[25] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. ”Deep residual learning for image recognition." in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 770-778. 2016

Prevalence of Diabetes Mellitus Among Human Immune Deficiency Virus-Positive Patients Under Anti-retroviral Attending in Rwanda, a Case Study of University Teaching Hospital of Butare

Venuste Kayinamura, V. Iyamuremye, A. Ngirabakunzi

Abstract**—** Anti-retroviral therapy (ART) for HIV patient can cause a deficiency in glucose metabolism by promoting insulin resistance, glucose intolerance, and diabetes, diabetes mellitus keep increasing among HIV-infected patients worldwide but there is limited data on levels of blood glucose and its relationship with antiretroviral drugs (ARVs) and HIV-infection worldwide, particularly in Rwanda. A convenient sampling strategy was used in this study and it involved 323 HIV patients (n=323). Patients who are HIV positive under ARVs were involved in this study. The patient's blood glucose was analyzed using an automated machine or glucometer (COBAS C 311). Data were analyzed using Microsoft Excel and SPSS V. 20.0 and presented in percentages. The highest diabetes mellitus prevalence was 93.33 % in people aged >40 years while the lowest diabetes mellitus prevalence was 6.67% in people aged between 21-and 40 years. The P-value was (0.021). Thus, there is a significant association between age and diabetes occurrence. The highest diabetes mellitus prevalence was 28.2% in patients under ART treatment for more than 10 years, 16.7% were <5years while 20% of patients were on ART treatment between 5-10 years. The P-value here is (0.03), thus the incidence of diabetes is associated with long-term ART use in HIV-infected patients. This study assessed the prevalence of diabetes among HIV-infected patients under ARVs attending the University Teaching Hospital of Butare (CHUB), it shows that the prevalence of diabetes is high in HIV-infected patients under ARTs. This study found no significant relationship between gender and diabetes mellitus growth. Therefore, regular assessment of diabetes mellitus especially among HIV-infected patients under ARVs is highly recommended to control other health issues caused by diabetes mellitus.

Keywords**—** anti-retroviral, diabetes mellitus, antiretroviral therapy, human immune deficiency virus.

Venuste Kayinamura, V. Iyamuremye, A. Ngirabakunzi are with the Catholic University Rwanda (CUR), Biomedical laboratory Science, Butare, Huye-Rwanda. PO.BOX 49 (e-mail: kayinamuravenuste1@gmail.com).

G. Nzeyimana is with the University Teaching Hospital of Butare (CHUB), Butare, Huye-Rwanda. PO.BOX.

L. Dusengemung is with the School of Mathematics and Natural Sciences, The Copperbelt University, Kitwe, Zambia.

Nigeria's Terrorists RehabIlitation and Reintegration Policy: A Victimological Perspective

Ujene Ikem Godspower

***Abstract*—** Acts of terror perpetrated either by state or non-state actors are considered a social ill and impugn on the collective well-being of the society. As such, there is the need for social reparations, which is meant to ensure the healing of the social wounds resulting from the atrocities committed by errant individuals under different guises. In order to ensure social closure and effectively repair the damages done by anomic behaviors, society must ensure that justice is served and those whose rights and privileges have been denied and battered are given the necessary succour they deserve. With regards to the ongoing terrorism in the Northeast, the moves to rehabilitate and reintegrate Boko Haram members have commenced with the establishment of Operation Safe Corridor,1 and a proposed bill for the establishment of "National Agency for the Education, Rehabilitation, De-radicalisation and Integration of Repentant Insurgents in Nigeria"2. All of which Nigerians have expressed mixed feelings about. Some argue that the endeavor is lacking in ethical decency and justice and totally insults human reasoning. Terrorism and counterterrorism in Nigeria have been enmeshed in gross human rights violations both by the military and the terrorists, and this raises the concern of Nigeria's ability to fairly and justiciably implement the deradicalization and reintegration efforts. On the other hand, there is the challenge of the community dwellers that are victims of terrorism and counterterrorism and their ability to forgive and welcome back their immediate-past tormentors even with the slightest sense of injustice in the process of terrorists reintegration and rehabilitation. With such efforts implemented in other climes, the Nigeria's case poses a unique challenge and commands keen interests by stakeholders and the international community due to the aforementioned reasons. It is therefore pertinent to assess the communities' level of involvement in the cycle of reintegration- hence, the objective of this paper. Methodologically as a part of my larger PhD thesis, this study intends to explore the three different local governments (Michika in Adamawa, Chibok in Borno, and Yunusari in Yobe), all based on the intensity of terrorists attacks. Twenty five in-depth interview will be conducted in the study locations above featuring religious leaders, Community (traditional) leaders, Internally displaced persons, CSOs management officials, and ex-Boko Haram insurgents who have been reintegrated. The data that will be generated from field work will be analyzed using the Nvivo-12 software package, which will help to code and create themes based on the study objectives. Furthermore, the data will be content-analyzed, employing verbatim quotations where necessary. Ethically, the study will take into consideration the basic ethical principles for research of this nature. It will strictly adhere to the principle of voluntary participation, anonymity, and confidentiality.

***Keywords*—** boko haram, reintegration, rehabilitation, terrorism, victimology.

Ikem Godspower Ujene is with the Achievers University, Owo, Ondo State, Nigeria, Nigeria (e-mail: profikechukwu@hotmail.com).

Using Action Research to Digitize Theses and Journal Articles at the Main Library, Sultan Qaboos University, Oman

Nabhan H. N. Al-Harrasi

***Abstract*—** Action Research (AR) plays an important role in improving the problematical situation. It is a process that enhances thinking and practise and bridges the gap between abstract and concrete thinking. Nowadays, AR as a methodology is wildly used to implement projects based on understanding the needs of owners, considering the organizational culture, meeting the requirements, encouraging partnership, representing different viewpoints, and building the project. This research describes the whole processes of digitizing Post-graduate theses and all articles published in 6 Journals at Sultan Qaboos University. AR implemented to respond to the university needs to enhance accessibilities to its information resources and make them available through the national repository. In order to prepare the action plan, the library administration met to discuss several points related to the proposed project, the most important of which are:

- Providing digitalization devices.
- Locating a specific part of the Library as a Digitization Unit.
- Choosing a team.
- Defining tasks.
- Implementing the proposed project and evaluating the whole processes.

***Keywords*—** action research, digitization, Theses, Journal articles, open access, Oman.

Nabhan Alharrasi is with the Sultan Qaboos University, Oman (e-mail: nabhan@squ.edu.om).

Agency Beyond Metaphysics of Subjectivity

Erik Kuravsky

Abstract— One of the problems with a post-structuralist account of agency is that it appears to reject the freedom of an acting subject, thus seeming to deny the very phenomenon of agency. However, this is only a problem if we think that human beings can be agents exclusively in terms of being subjects, that is, if we think agency subjectively. Indeed, we tend to understand traditional theories of human freedom (e.g., Plato's or Kant's) in terms of a peculiar ability of the subject. The paper suggests to de-subjectivize agency with the help of Heidegger's later thought. To do it, ir argues that classical theories of agency may indeed be interpreted as subject-oriented (sometimes even by their authors), but do not have to be read as such. Namely, the claim is that what makes agency what it is, what is essential in agency, is not its belonginess to a subject, but its ontological configuration. We may say that agency "happens," and that there is a very specific ontological characteristics to this happening. The argument of the paper is that we can find these characteristic in the classical accounts of agency and that these characteristics are sufficient to distinguish human freedom from other natural phenomena. In particular, it offers to think agency not as one of human characteristics, but as an ontological event in which human beings take part. Namely, agency is a (non-human) characteristic of the different modes in which the experienceable existence of beings is determined by Being. To be an agent then is to participate in such ontological determination. What enables this participation is the ways human beings non-thematically understand the ontological difference. For example, for Plato, one acts freely only if one is led by an idea of the good, while for Kant the imperative for free action is categorial. The agency of an agent is thus dependent on the differentiation between ideas/categories and beings met in experience – one is "free" from contingent sensibility in terms of what is different from it ontologically. In this light, modern dependence on subjectivity is evident in the fact that the ontological difference is thought as belonging to one's thinking, consciousness etc. That is, it is taken subjectively. A non-subjective account of agency, on the other hand, requires thinking this difference as belonging to Being itself, and thinking human beings as a medium within which occurs the non-human force of ontological differentiation.

Keywords— Heidegger, freedom, agency, poststructuralism.

Erik Kuravsky is with the Tel Aviv University, Israel (e-mail: erikkuravsky@gmail.com).

The Development of Da'wah and Challenges to The Issue of Mualaf in Malaysia

Hailan Salamun
Universiti Malaysia Terengganu (UMT), Terengganu, Malaysia
hailan@umt.edu.my

Firdaus Khairi Abdul Kadir
Universiti Malaysia Terengganu (UMT), Terengganu, Malaysia
firdauskhairi@umt.edu.my

Nazihah Rusli
Universiti Malaysia Terengganu (UMT), Terengganu, Malaysia
nazihahrusli95@gmail.com

ABSTRACT

Malaysian society is a multi -religious and multi -cultural society that makes this country often emulated by the international community. The current situation of the development of Islamic da'wah is a phenomenon that shows the increase in the number of Muslims proves that Islam has become a universal religion, and people are beginning to recognize and accept the truth of Islam. Even so, the issues related to converts have never been resolved as it is a difficult matter to deal with. Despite that, the phenomenon of mualaf returning to their former religion had concerned the authorities as reported by newspapers and social media. In general, this study is conducted to deepen and describe the issues and challenges of the Mualafs which are said to be increasingly complex among the Malaysian community. This study is to identify the causes of converts returning to the previous religion based on cases of apostasy that occurred in Malaysia. This qualitative document analysis technique uses secondary data in attaining the information related to the study. The findings of the study indicate that the main factor for most converts to accept Islam is through marriage, the encouragement of friends and some research done by them. However, Mualafs who face challenges and problems in life will usually lead them to negligence and indifference to the teachings of Islam. Thus, the issues and challenges of Mualaf, namely the lack of iman (‘aqidah) in Islam, insufficient knowledge of Islam, lack of religious commitment, failure in marriage and biological family problems. On the other hand, this study will assist the preacher and *da’wah* institution to comprehend more about issues related to mualaf and suggest a more practical method of *da’wah* for them. By that, mualaf would be able to endure the urge to commit murtad after facing the hardships in their life, and then the murtad cases in Malaysia simultaneously could be put to an end.

Keywords: Mualaf, Murtad (apostasy), Murtad case, Faith (‘aqidah), Da’wah

Fields: Philosophy and Civilization

Sub-Field: Islamic Studies/Da'wah

INTRODUCTION

A person who has just converted to Islam from a previous religion is known as a 'Mualaf'. The use of the term 'Mualaf' has become commonplace in society. There are also 'Mualafs' who are more comfortable being called 'new brothers' or 'Muslim brothers'. They felt the call was more friendly and did not indicate too much distance between fellow believers. However, there are a number of Mualaf who face economic problems, do not get a perfect Islamic education and lack of involvement in religious classes and get the perfect guidance can lead to not practicing the teachings of Islam in their lives. These people who end up easily stuck with the problem of returning to the original religion or called 'murtad'.

In Arabic, murtad comes from the word *riddah*. In terms of language, *riddah* means the return of something otherwise (al-Razi, n.d, p.101). In Islamic (*syarak*) terms, murtad is defined as renounce Islam and return to disbelief (*kufur*) and detach themselves from Islam (al-Husni, 2016, p.637). According to Kamus Dewan Edisi Keempat, murtad means a person who turned away from Islam, whether through act or word, or intention; they are in disbelief (*kufur*), rebellious and unfaithful towards Allah SWT (The Almighty).

In Islam, murtad (apostasy) is categorized as a crime and it is clearly stated in the Qur'an surah al-Baqarah, verse 217, which says,

> "*…while whoever of you turns away from his faith and dies an infidel, such people are those whose deeds will go to waste in this world and the Hereafter, and they are people of the Fire. They shall be there forever*" (translated by Mufti Taqi Usmani).

Based on the rough translation above, Muslims must avoid apostasy at all costs. It is because apostasy falls in the most extreme category of sin that will not be forgiven by God (Allah SWT).

Mualaf and current issues

In Malaysia, when there is an apostasy issue amongst Muslims, they go to the Shariah Court to apply the declaration of apostasy and each state will has different laws to deal with the issue. This is because the laws related to religion or *syariah* included murtad (apostasy) will be discussed at state government (Suariza, Raudah & Yusmini, 2018). Despite that, the phenomenon of Muslims had applied to leave Islam especially mualaf apparently because of various factors that forced the authorities to deal with this issue regularly.

Malaysiakini (2011) notified that Mufti of Selangor, Datuk Tamyes Abdul Wahid stated many mualaf had applied to leave Islam after conversion. He mentioned some of them convert to Islam only to get married but then the failure of the marriage caused them to return to their former religion. He also pointed out the role of parents or society in guiding mualaf in order for them commit to Islamic teachings. Besides, Suariza, et al., (2018) stated in their study based on the statistic given by the department of prime minister, Shariah Court had received 863 cases from the year 2000 until 2010 to change the religion and mostly the application comes from mualaf. They pointed out the problem that caused mualaf to commit murtad is because of challenges and sacrifices faced by them are coming from their family or the surrounding community.

In the meantime, *da'wah* in a plural society is not easy since it involves physical, mental and emotional. Nowadays, there are numerous activities and programs in helping mualaf to motivate them with the aim they would stay committed with religious commitment. However, issues related to mualaf are still problematic and in need of efficient solutions. The fact proved by Noreha, Asmawati & Fathiyah (2019) stressed in their study about the inconsistency of *da'wah* activities to mualaf. They stated the various programs conducted by different organizations indirectly leading mualaf in confusion to learn about Islam also, the unorganized management of mualaf still existed. In such a way, mualaf need continuous assistance and guidance to strengthen their understanding and belief in Islam. Casmini (2020), stated in her study, mualaf had to adapt to their new religion since they had to experience in shifting of their identity, values, and behaviour that affects their social life. So, they would face numerous difficulties along the journey of seeking their new religious identity after converting to Islam.

Furthermore, Fatimah, Nadhirah, Khatijah, and Hajar (2018), mentioned in their study that new life after conversion demanded mualaf to abide by Islamic rules and regulations, a commitment to practice Islamic values and build strong faith in Islam. They need to face all problems and difficulties in their life with patience and persistence, either internal or external factors. Moreover, the challenge of understanding Islam as a religion had left mualaf with no choice but to seek nonstop guidance from others until there is an ease in getting through their religious identity simultaneously have resilience in facing new hardships after being a Muslim.

Thus, this study emphasizes mualaf as a complex issue that has never ending-problems and is in dire need of a more practical approach to deliver *da'wah*. It is mainly related to the need for them to have strong and steady faith (*'aqidah*) in Islam. Researcher wants to seek the reasons that caused mualaf to return to their former religion and shortcoming the effort taken by the authorities in dealing with this issue. It will leave a significant impact on the implementation of *da'wah* in Malaysia especially involved with mualaf. Therefore, the objective of this study is to identify the reasons that caused mualaf to return to their former religion after converting to Islam based on the murtad cases that happened in Malaysia.

LITERATURE REVIEW

Mualaf is the most important asset in *da'wah Islamiyyah* because they are the person who is tamed their heart to receive a guide (*nur*) from Allah SWT to embrace Islam as their new religion. Nonetheless, the murtad cases happened among them are not an exceptional situation since they had to experience shifting between different culture and way of life from their previous religion. The preparation to survive all of the difficulties in life after converting to Islam is essential in order to fulfil the religious commitment in case, they lost their faith in Islam.

Titian and Rudi (2015) stated in their study there are five dimensions in conceptual of religious commitment which are: knowledge, belief, practice, experience and consequence. They believe those five dimensions can identify the level of commitment of mualaf towards religious teaching. This is because without commitment towards their religion, it will be difficult for them to do well with their religious life. Therefore, to fulfil the expectation towards the religion, mualaf should learn a lot to master Islamic teaching. Consequently, mualaf would be put at ease and genuinely practicing the religious teaching and at the same time, they would stay to have faith in Islam.

Nevertheless, the murtad cases that happened among mualaf in Malaysia could happened for various factors. The possibility of main concern for them is the challenges to adapt to new religious identity as they had to go through in clash of cultures and different lifestyles from their former religion. Sometimes, murtad cases occur due to the lack of attention given to them, life challenges and poor management in dealing with the issue related to them (Nizam, Aishah & Suhaila, 2013). In their study, they addressed murtad cases happened as a result of hardships in life after conversion to Islam without described more the details on why mualaf choose to renounce Islam.

After an individual had liberated himself/herself from Islam and return to disbelief (*kufur*), they had to go through counselling and be given advice so that they would repent as much as they need. According to the study by Zaleha, Sarah & Faisal (2016) the duration of repent (*taubat)* has two opinions either three days and three nights or 20 days. If they had realized their mistake (sin), their *taubat* accepted through the process; pronounce the Shahadah again and admit their sin. Based on the statement above, it showed that faith (*'aqidah*) in Islam is the most vital for mualaf to survive with their new religious identity. In addition, Yunus, Samsuddin & Misnan (2017) quoted faith education as notably the most necessary aspect in teaching mualaf about Islam as leaving them with weak faith would be one of the major reasons which lead them to go back to their previous religion. They claimed the level of guidance and knowledge about Islam has a significant correlation to trigger off symptoms of murtad (apostasy) among mualaf.

The fact supported by Kamal, Hussin & Rosni (2020) cited in their research *da'wah* Prophet Muhammad SAW has priority in which their main mission is teaching *'aqidah* since *'aqidah* is the basic matter in Islam. Even there are three important elements in Islam; *'aqidah, syariah* and *akhlak,* the most vital thing is *'aqidah* to develop s*yariah* and *akhlak.* For this reason, mualaf tend to return to their former religion because of the weakness of *'aqidah* as a result of lack of guidance and challenges in practicing Islamic teaching leading to neglect of religious commitment.

On top of that, in one study, Buhar, Syukri & Zawiyah (2013) clearly stated some problems faced by mualaf after conversion, for example, lack of knowledge, realization, confirmation, or confused acceptance of the religion or the sociological complexities or the negative impact of surroundings. They explained mualaf experienced various conflicting feelings towards Islam on negative insights as adopting a new religion caused them to traumatic inner conflicts because of different traditions of old religious faith. In this way, they believe this issue should be resolved using the real *tauhidic* (monotheistic) message of al-Qur'anic and al-Sunnah, an approach in which the purpose is to go after the goodness, prevent harm, and decline the evil; *al-masolih al-dharuriyyah* in *Maqasid Syariah.* This is crucial to the point it is essential and cannot be avoided for the sake of humans' religion and life. In the Qur'an surah al-Hadid, verse 25, which says,

> "*We sent Our messengers with clear signs, the Scripture and the Balance so that people could uphold justice: We also sent iron, with its mighty strength and many uses for mankind, so that God could mark out those who would help Him and His messengers though they cannot see Him. Truly God is Powerful, Almighty*" (translated by Abdul Haleem).

Also, surah al-Ma'idah, verse 8, which says,

> *"O you who have believed, be persistently standing firm for Allah, witnesses in justice, and do not let the hatred of a people prevent you from being just. Be just; that is nearer to righteousness. And fear Allah; indeed, Allah is [fully] Aware of what you do"* (translated by Saheeh International).

Accordingly, the translation clearly described the importance to achieve a motive of objective of *syariah* that discusses such measures and principles in protecting the right of people without discrimination between skin colour, race or nation. Hence, it means the right of mualaf is no different from born-Muslim, which need to be shielded no matter what and that is why these problems faced by them have to be settle thoughtfully.

Consequently, even though there are many efforts taken by the researcher to discuss and conduct the study related to issues of mualaf and murtad cases that happened among mualaf, it is still not enough given this issue is complicated and has numerous problems. It would be enough reason for the researcher to seek a more efficient and practical approach in dealing with this concern. In such a manner, the authorities should be working together to propose a more practical and consistent method in approaching mualaf since they need continuous guidance to understand Islam and fulfilling religious commitment, as well as mualaf, can resist the urge to leave Islam.

RESEARCH DESIGN

This study is a qualitative design using the document analysis technique. Data collection for this approach was reviewed and analysed from secondary data such as information from the internet, articles, journals, newspapers, and so on. This study also was analysed the cases of murtad that happened in Malaysia and then will identify the main cause of why some mualaf choose to return to their former religion and commit murtad. All findings of the study will be summarized and presented in the findings and discussion.

RESULTS AND DISCUSSION

Mualaf indeed had gone through various challenges and problems in their life after converting to Islam. They lack resistance in dealing with their hardships to survive with new religious life. The findings of this study based on the murtad cases that happened in Malaysia showed many reasons that caused mualaf to return to their former religion. Some of them are:

Lack of faith (*'aqidah*) in Islam

One of the main reasons that caused mualaf to leave Islam. If this the most basic thing in Islam was neglected, it does no wonder mualaf felt lost and confused to get the hang of knowledge about Islam. After all, they need to adapt to their new religious life since their culture and way of life from their previous religion is much different. Then, they choose to renounce Islam after do not have enough faith in Islam; Allah SWT (The Almighty) is the Only that Muslims should believe as He is the one who plans everything but, they still choose to leave Islam cause of this weakness of faith (*'aqidah*).

Insufficient knowledge about Islam

After converting to Islam, the first thing mualaf should do is to learn everything about Islam. Being a mualaf at least they need to master the basic matters in Islam such as fardhu ain, prayer (*solah*), *Rukun* Islam, *Rukun* Iman, and the way of purification (*hadas*). They would not choose to go back to their former religion if they acquire this knowledge (*'ilmu*).

Lack of religious commitment

As stated by Titian and Rudi (2015) above, there are five dimensions in conceptual of religious commitment which is: knowledge, belief, practice, experience, and consequence. If mualaf had adept these five dimensions there is no way they would commit murtad since they had a chance to fulfil the expectation of the religion. However, based on some murtad cases that happened among mualaf, they failed to obligate religious commitment and still practice their old religious' habit.

Failure in marriage

Based on the report or news about murtad cases among mualaf, a big percentage of it happened related to marriage. Many mualaf converts to Islam because they want to get married to Muslim but, some of them had gone through a few problems like got divorced, being left with no reason by their Muslim partner, and being broken of engagement. There are some cases where mualaf got married to a Muslim partner, but their partner does not guide them with Islamic teaching and allowed mualaf to practice former religion's habit. In that way, it could trigger them to return to their former religion since they cannot see the truth and beauty of Islam as well as got lost and confused about religious teaching.

Trouble from the birth family

There are cases of mualaf got abandoned from their family institution and boycotted by friends just because they embraced Islam. They received threats and harm as their family and friend cannot agree with them being Muslim. Moreover, there is also a murtad case where mualaf was raised as a Hindu by grandma even though the parents converted. This is an example of dilemma and confusion experienced by mualaf which leads them to not have a chance in learning and practicing Islamic teaching. Then, they were forced to commit murtad since that is the only option as they think based on their situation.

CONCLUSION

Upon embracement of Islam, mualaf had experienced many changes in their life, namely psychological or social factors. Issues like social changes, daily routines, cultural integration, familial relationship, and emotional changes would raise problems among them if these concerns do not be tackled wisely. Therefore, appropriate and effective method should be taken to help mualaf to motivate them along with their journey as Muslim since they need to survive with a new religious identity. Inconsistency in *da'wah* activities caused mualaf left with confusion and then the feeling of loss would create uneasiness in them to fulfil the religious commitment. Indeed, mualaf had encountered challenges and problems of which they cannot avoid. The authorities including *da'wah* institutions and the preachers should lend a hand and work together in seeking the right answer for these issues. As discussed above, the issues of mualaf are complicated and it begs for a more practical approach to be implemented. The most realistic approach should be the one that included all the aspects in living such as *Maqasid Syariah* that has five principles, protection of religion (*al-din*), protection of life (*al-nafs*), protection of lineage/dignity (*al-nasb*), protection of intellect (*al-'aql*) and protection of property (*al-mal*). For this reason, the objective of Islamic law in which it aims to pursue goodness and benefit of people both in this world and the Hereafter could be accomplished as well as Islam will be a reality in life when it is practiced genuinely.

ACKNOWLEDGEMENT

This study is a part of fundamental research grant scheme (FRGS/1/2020/SS0/UMT/02/10) under MOHE (KPT) Malaysia conducted together with other researchers from UMT and UniSZA.

REFERENCES

Ahmad Yunus Kasim, Samsuddin Abdul Hamid & Misnan Jemali. (2017). Pengajaran Akidah dalam Kalangan *Mualaf* di Institut Dakwah Islamiyah PERKIM. *Jurnal Perspektif: Special Issue 1*, 89-100.

Al-Husni, Taqi al-Din Abu Bakr bin Muhammad. (2016). *Kifayat al-Akhyar fi Hil Ghayat al-Ikhtisar*. Jeddah: Dar al-Minhaj. Retrieved from https://archive.org/details/FP163503/page/n2/mode/1up?q=riddah

Al-Razi, Muhammad bin Abi Bakr bin Abd al-Qadir. (n.d.). *Mukhtar al-Sihah*. Beirut: Maktabah Lubnan. Retrieved from https://archive.org/details/waq8477/8477/page/n100/mode/1up

Casmini. (2020). Analysis of Muallaf 'Aisyiyah Da'wah Strategy. *Ilmu Dakwah: Academic Journal for Homiletic Studies* (1) 151-166. DOI:10.15575/idajhs.v14i1.9238.

Kamal Azmi Abd Rahman, Mohd Noor Hussin & Rosni Wazir. (2020). Keberkesanan Pengajian Akidah bagi Mualaf: Analisis Keperluan. Proceeding from conference: *International Seminar on Muallaf* (ICOM 2019). Retrieved from https://www.researchgate.net/publication/340573369_Keberkesanan_Pengajian_Akidah_Bagi_Mualaf_Analisis_Keperluan

Kamus Dewan Edisi Keempat. (2017). *Dewan Bahasa dan Pustaka*. Retrieved from https://prpm.dbp.gov.my/Cari1?keyword=murtad

Mohd Nizam Sahad, Siti Aishah Chu Abdullah & Suhaila Abdullah. (2013). Malaysian News Report on Muslim Converts' Issues: A Study on *Malaysiakini. International Journal of Humanities and Social Science*, Vol.3, No.13, 219-230.

Norain Saleh. (2011, June 17). Mufti: Ramai mualaf yang pohon murtad. *Malaysiakini*. Retrieved from https://www.malaysiakini.com/news/167179

Noreha Che Abah, Asmawati Suhid & Fathiyah Mohd Fakhruddin. (2019). Isu dan Cabaran Saudara Baru di Malaysia: Satu Tinjauan Awal. *Jurnal AL-ANWAR*, volume 8, No.2, 1-13.

Quran (al-Baqarah) 2:217

Quran (al-Hadid) 57:25

Quran (al-Ma'idah) 5:8

Razaleigh Muhamat Kawangit. (2016). Pembangunan Dakwah Muallaf Di Malaysia: Cabaran Dalam Masyarakat. Dalam conference 1st International Keynote Speech on Mualaf Development and Empowerment (DE MUALAF). Dimuat turun daripada (19) (PDF) PEMBANGUNAN DAKWAH MUALLAF DI MALAYSIA: CABARAN DALAM MASYARAKAT (researchgate.net) pada 5 Disember 2021.

Sayyid Buhar Musal Kassim, Mohd Syukri Yeoh Abdullah & Zawiyah Baba. (2013). A Survey of Problems Faced by Converts to Islam in Malaysia. *Journal of Social Sciences and Humanities*, Vol.8, No.1, 085-097.

Siti Fatimah Salleh, Nadhirah Nordin, Siti Khatijah Ismail & Siti Hajar Mohamad Yusoff. (2018). Cabaran dan Implikasi Pengurusan Dokumentasi Saudara Baru. Proceedings from International Seminar on *al-Quran in Contemporary Society 2018* organized by University Sultan Zainal Abidin (UniSZA) eISBN 978-967-0899-96-1.

Siti Zaleha Ibrahim, Nur Sarah Tajul Urus & Dr. Mohd Faisal Mohamed. (2016). Pertukaran Agama dan Kesannya Terhadap Komuniti: Satu Sorotan Terhadap Kes-kes Murtad dan Masuk Islam di Malaysia. *Journal of Social Sciences and Humanities*, Special Issue 3, 204-213, ISSN: 1823-884.

Suariza@Hidayah Muhammad, Nor Raudah Hj. Siren & Yusmini Md Yusoff. (2018). Faktor Permohonan Isytihar Murtad dalam Kalangan Mualaf di Selangor. *Jurnal Usuluddin* 46 (2), 123-146.

Titian Hakiki & Rudi Cahyono. (2015). Komitmen Beragama pada Muallaf (Studi Kasus pada Muallaf Usia Dewasa). *Jurnal Psikologi Klinis dan Kesehatan Mental*, Vol.4 No.1, 20-28.

The Role of Human Beings as Caliphs in Preserving Nature

Firdaus Khairi Abdul Kadir, Nazihah Rusli, Noor Aisyah Abdul Aziz

***Abstract*—** Islam is a comprehensive religion encompassing all aspects of society's life such as social, economic, political, cultural and environmental. The environment is part of the manifestation of God's greatness which has pearls of wisdom, bestowed upon human beings to make them realize that everything is in the hands of God (Allah SWT). However, the equilibrium of nature could be disturbed from the excessive exploitation by humans' hands. As a caliph on this earth, it is the responsibility of human beings to look after the environment proactively. Besides, Islam calls for the execution of accountable development and respecting the principles of sustainability. Therefore, this study focuses on the role of human beings as caliphs on this earth who are responsible for nature and their acts in conserving and preserving the environment. This study also used the research method of the survey library.

***Keywords*—** environment, human beings, caliph, *tauhid* and Allah SWT.

I. Introduction

THE entire universe falls under the control and direction of Allah SWT from the smallest things to the largest of His creatures whether it is can be seen by human beings or not. Muslims believe in the existence of the invisible (*ghaib*) world which generally included other than angels, devils, the hereafter, heaven, hell and demon (*jinn*). The same goes with nature that is visible with the naked eyes such as water, plants, animals and so on that existed around human beings. Nevertheless, all of these are the creations of Allah SWT that made it with perfection in which nature and its contents have their place and role. Additionally, whereby just as Allah SWT created the gases in the air, the process of photosynthesis takes place, the existence of human beings, animals, plants and the circulation of the moon, sun and star in the universe beyond earth's atmosphere, for the sake of nature's equilibrium.

According to [1], the term environment refers to the whole of external factors and conditions that affect living organisms, for instance, air, water, light, animal, humans, the sun, and others. Meanwhile, the dictionary of Idris al-Marbawi, [2] defined nature with the meaning of "ما سِوى الله", which means "anything other than Allah SWT". In this way, it can be concluded that the definition of nature from an Islamic perspective is all the creatures created by Allah SWT both visible and invisible things. Those creatures can be categorized into four; first, the creatures that can be seen with the sense of sight such as stone, wood, water, and others. Second, the creature which is cannot be seen with naked eyes but can be seen with a man-made device like a microscope, for example, atoms and bacteria. The third is the beings which can be felt physically but cannot be seen with the eye like wind, air, and electric current. Lastly, the creatures who are their existence can be felt instinctively, however, not capable to be seen with any of the senses or any tools of humans' creation such as angels, devils, and demons (*jinn*).

II. Review of Literature

The Book of Allah SWT (al-Quran) highlighted that human beings as a special position among other of God's (Allah SWT) creations on earth since they are endowed with intellect ('aql) compared to other creatures. For this reason, they have been chosen as the representative in taking care of God's creation on this earth. This is stated by Allah SWT in surah al-Isra, verse 70, which says [3],

۞وَلَقَدْ كَرَّمْنَا بَنِىٓ ءَادَمَ وَحَمَلْنَـٰهُمْ فِى ٱلْبَرِّ وَٱلْبَحْرِ وَرَزَقْنَـٰهُم مِّنَ ٱلطَّيِّبَـٰتِ
وَفَضَّلْنَـٰهُمْ عَلَىٰ كَثِيرٍ مِّمَّنْ خَلَقْنَا تَفْضِيلًا(٧٠□

> "Indeed, We have dignified the children of Adam, carried them on land and sea, granted them good and lawful provisions, and privileged them far above many of Our creatures" (translated by Dr Mustafa Khattab).

Accordingly, as a caliph on this earth, human beings should obligate and act based on the decree of Allah SWT. They need to realize and be aware that even though Allah SWT made human beings the best of His creation, they also would become the lowest and dirtiest creature if they were unfaithful towards Allah SWT and His Messenger. It is proved in surah at-Tin, verse 4 until 6, which says [4],

لَقَدْ خَلَقْنَا ٱلْإِنسَٰنَ فِيٓ أَحْسَنِ تَقْوِيمٍ(٤□ ثُمَّ رَدَدْنَٰهُ أَسْفَلَ سَٰفِلِينَ(٥□ إِلَّا ٱلَّذِينَ ءَامَنُواْ
وَعَمِلُواْ ٱلصَّٰلِحَٰتِ فَلَهُمْ أَجْرٌ غَيْرُ مَمْنُونٍ(٦□

> "We have certainly created man in the best of stature; (4) Then We return him to the lowest of the low, (5) Except for those who believe and do righteous deeds, for they will have a reward uninterrupted" (6) (translated by Saheeh International).

Other than that, a human who has a strong belief in tauhidiyyah (oneness of Allah SWT) would be free from any challenges in fulfilling the syariah duties. They should be aware of their responsibility as soon as Allah SWT has entrusted them with it and complete themselves with sufficient knowledge to

Firdaus Khairi Abdul Kadir is with the Centre for Fundamental and Continuing Education, Universiti Malaysia Terengganu (UMT), Malaysia (corresponding author, e-mail: firdauskhairi@umt.edu.my).

Nazihah Rusli is with the Centre for Fundamental and Continuing Education, Universiti Malaysia Terengganu (UMT), Malaysia.

Noor Aisyah Abdul Aziz is with the Centre for Fundamental and Continuing Education, Universiti Malaysia Terengganu (UMT), Malaysia.

comprehend the concept of the environment that would be governed. Al-Quran had mentioned in surah al-A'raf, verse 74, which says [5],

وَٱذۡكُرُوٓاْ إِذۡ جَعَلَكُمۡ خُلَفَآءَ مِنۢ بَعۡدِ عَادٖ وَبَوَّأَكُمۡ فِي ٱلۡأَرۡضِ تَتَّخِذُونَ مِن سُهُولِهَا
قُصُورٗا وَتَنۡحِتُونَ ٱلۡجِبَالَ بُيُوتٗاۖ فَٱذۡكُرُوٓاْ ءَالَآءَ ٱللَّهِ وَلَا تَعۡثَوۡاْ فِي ٱلۡأَرۡضِ مُفۡسِدِينَ(٧٤□

"And remember when He made you successors after the ʿAad and settled you in the land, [and] you take for yourselves palaces from its plains and carve from the mountains, homes. Then remember the favors of Allah and do not commit abuse on the earth, spreading corruption" (translated by Saheeh International).

Meanwhile, as sociable beings, human beings have the same ecology and biology for the continuation in life same as other beings. Also, the creature that tends to be completely dependent on Allah SWT should have a feeling of gratitude and a sense of responsibility to preserve the nature of God's (Allah) creation. This message is recorded in the Quran surah an-Naml, verse 31, which says [6],

أَلَّا تَعۡلُواْ عَلَيَّ وَأۡتُونِى مُسۡلِمِينَ(٣١□

"Be not haughty with me but come to me in submission [as Muslims]" (translated by Saheeh International).

Moreover, the act of things that forbids by Allah SWT to the point of destroying well-created creations seems like showing the discourteous that is means disrespect towards Him; the Supreme. Human beings need to realize that everything that exists in this universe belongs to Allah SWT solely, not their explicit possession even though they are given the freedom to use natural resources. In addition, they cannot misuse these natural resources beyond their expectations and needs. Allah SWT reminded of this matter in surah al-A'raf, verse 85, which says [7],

..وَلَا تُفۡسِدُواْ فِي ٱلۡأَرۡضِ بَعۡدَ إِصۡلَٰحِهَاۚ ذَٰلِكُمۡ خَيۡرٞ لَّكُمۡ إِن كُنتُم مُّؤۡمِنِينَ(٨٥□

"...and do not defraud people of their property, nor spread corruption in the land after it has been set in order. This is for your own good, if you are 'truly' believer" (translated by Dr Mustafa Khattab).

The Quran had emphasized many times the role of human beings as caliphs in general, and Muslims in particular, to bear the responsibility as guardians of the environment, and then that every action will be questioned on the Day of Judgement. Despite that, the disruption that happened nowadays that caused equilibrium in the ecosystem had proved there was too much environmental damage already. Even though for thousand years ago Islam had pointed out the importance of plants in preserving the environment and at the same time, able to reduce the effects of climate change. Consequently, Prophet Muhammad SAW had always forbidden his companions to destroy trees during the war as well as stressed the act of planting trees. The Prophet SAW had said in a hadith, which says, 'there is no Muslim who plants a tree or sow seeds, and then a bird or a person or an animal eats from it but, is considered a charitable gift for him' (Bukhari Hadith).

Furthermore, during the war also, Prophet Muhammad SAW distinctly prohibit the destruction of trees and plants because of its advantage to be used as shelter to the troops. It said in the previous hadith, the benefits of plants or trees are given to others unintentionally.

On the other hand, the climate change that occurs nowadays is one of the shreds of evidence in which the ones who are responsible for it had failed to carry out their duty. Industrialization had led to habitat destruction dramatically whereby forests are cut down for the wood and ecosystem were disturbed to create roads, strip mines, and gravel pits. Besides, tearing down these habitats could affect the local ecosystem and lead to the extinction of plants and animals since the species cannot migrate and adapt to the new environment. Human beings have often betrayed the warning and reminder about how important to keep their eyes open on this nature by committing irresponsible acts. This situation had stated in the Quran surah ar-Rum, verse 41, which says [8],

ظَهَرَ ٱلۡفَسَادُ فِي ٱلۡبَرِّ وَٱلۡبَحۡرِ بِمَا كَسَبَتۡ أَيۡدِي ٱلنَّاسِ لِيُذِيقَهُم بَعۡضَ ٱلَّذِي عَمِلُواْ
لَعَلَّهُمۡ يَرۡجِعُونَ(٤١□

"Corruption has appeared throughout the land and sea by [reason of] what the hands of people have earned so He [i.e., Allah] may let them taste part of [the consequence of] what they have done that perhaps they will return [to righteousness]" (translated by Saheeh International).

In the meantime, the combination of the concepts of tauhid (oneness of Allah SWT), khilafah (representative), and trust in the sustainable management of natural resources is supported by the Islamic perspective on environmental conservation [9]. The preservation of the environment should be understood as a religious order and every individual must be taken care of it. Then, religious worship ('ibadat) which is done because Allah SWT solely includes effort in the management and conservation of the environment is considered as an act of good deeds and dignified to the point every good effort would be rewarded by Him. Meanwhile, those actions that are contrary to the religious teaching not only cause damage to the environment but also certainly would face His wrath. The fact is, the environmental crisis that is happening today is due to human's greed and their failure to fulfil their trust as the guardians of nature. According to [10], the fade of responsible attitude towards environmental care is because of the lack of Islamic values' appreciation related to the environmental management based on al-Quran and as-Sunnah. The current modernization has witnessed the extreme action of humans by damaging the creation of God for the sake of profit alone [11]. Therefore, it is clear that preserving the sustainability of nature is the main challenge faced by human beings.

More than that, al-Quran which was revealed to Prophet Muhammad SAW is a complement to the Islamic law. Henceforth, Islam arises amid the world community to be the best solution for people (ummah) in this world. Human beings should put their effort into treating and preventing environmental damage by returning to the Islamic teaching as a perfect guide in their life in line with Maqasid Syariah (purpose of legislation). They can perform the sustainability of the environment with the guidance of Islam based on al-Quran. This matter is following what said by Prophet Muhammad SAW narrated by Abdullah bin Abbas in the book Fathu Barri which says,

الإسلامُ يعلو ، ولا يُعلى عليه

Meaning: Islam is the highest religion, nothing more than that.

Aside from that, [12] explained the three main areas that reflect the ethics of sustainable development; firstly, creating the equilibrium of the environment by appreciating its components; secondly, the value of environmental management that is focused on the human beings as the sole representative to look after the environment; and the third, is the plan for the environment to protect the environment.

Next, the advantage of science and the role of the scientist in the development of new technology for the sake of protecting the environment also needed to be respected. Green technology is seen to be able to help minimize the negative effect on human activities as well as contribute to the Islamic civilization [13]. Islam has understood the importance of trees in protecting the environment and reducing the effect of climate change since thousands of years ago, and that is why Islam has outlined the rules regarding the aspect of environmental sustainability.

III. Methodology

The primary methodology used in this writing is a library research design that uses the most relevant information from the secondary source. Data collection involved in this study is from books, Quran, hadith, articles, reports, and scholarly research; published both printed and online.

IV. Results and Discussions

The caliph plays the primary role during the creation of human beings. Caliph (khalifah) comes from the word khalafa (خلف) which means to follow or comes after him, or the most accurate term is 'substitute'. The concept of the human being as a caliph on this earth can be summarized into five main characteristics; 1) representative (khilafah), 2) trust, 3) leadership (qiyadah/siyadah), 4) religious worship (ibadah/ubudiyyah), 5) trial (ibtila') [10]. Besides, humans should be fully responsible to manage, govern and protect the earth. However, humans are often betraying the warning and reminder on how important to look after nature by committing irresponsible acts. As caliphs of Allah SWT, human beings should manage the natural resources in an orderly manner. Despite that, the Quran itself has warned against excessive waste and exploitation being committed [14].

On the other hand, according to reference [15], regarding the environment, it is stated that nature is the gift from Allah SWT to all beings (makhluk) who inhabit it. Moreover, for Muslims, the appreciation towards the environment is better if they explore the various message entrusted by Allah SWT through the Quran. The Quran has described from the Islamic view the significance of nature as stated in surah az-Zumar, verse 21, which says [16],

أَلَمۡ تَرَ أَنَّ ٱللَّهَ أَنزَلَ مِنَ ٱلسَّمَآءِ مَآءٗ فَسَلَكَهُۥ يَنَٰبِيعَ فِي ٱلۡأَرۡضِ ثُمَّ يُخۡرِجُ بِهِۦ زَرۡعٗا مُّخۡتَلِفًا أَلۡوَٰنُهُۥ ثُمَّ يَهِيجُ فَتَرَىٰهُ مُصۡفَرّٗا ثُمَّ يَجۡعَلُهُۥ حُطَٰمًاۚ إِنَّ فِي ذَٰلِكَ لَذِكۡرَىٰ لِأُوْلِي ٱلۡأَلۡبَٰبِ(٢١□

"Do you not see that Allah sends down rain from the sky and makes it flow as springs [and rivers] in the earth; then He produces thereby crops of varying colors; then they dry and you see them turned yellow; then He makes them [scattered] debris. Indeed, that is a reminder for those of understanding" (translated by Saheeh International).

Accordingly, the environment is inseparable from one's faith in God with the manifestation that it can be seen from human behaviour, by emphasizing morality (akhlak) as a core of the relationship between humans, nature, and God [17]. These relationships are needed to be stressed comprehensively to maintain and preserve sustainable development. Islam teaches its ummah, they are encouraged to use everything that is created in this universe properly and do not treat it carelessly which could lead to disaster in the future. However, behind those favors, lurks a test from Allah SWT to measure to which extent the nature of human trust to live in a way that pleases Allah SWT and their efforts to maintain the harmonious environment.

On top of that, know that Allah SWT has appointed human beings as caliphs on this earth so that they are responsible to maintain harmonious nature and look after the environment proactively. As well, the whole earth has been created to be a place of worship, clean and holy. For this reason, it is an obligation to human beings to preserve the natural resources wisely and thoughtfully considered that Allah SWT had given them exclusively the right to use and utilize its resources. Therefore, Allah SWT has created natural resources as a necessity to human beings not for them to be arrogant or feel proud yet, for the aim to human beings know the meaning of gratitude as well as able to be humbler in their selves (be more piety, taqwa).

V.Conclusion

Indeed, this beautiful universe has given many benefits to the alive beings as well as the dead. For the living, the benefits are very clear as proof of the existence of God (Allah SWT), and human beings subsequently would be worshipped themselves to Allah SWT in addition to seeking worldly reward as the reservation for the hereafter. Meanwhile, for the dead, the earth has made its land continue to be fertile for the others being in a way they could carry on with their life. Thus, all of those events are proof of His power over everything in this world. If human beings observe further, they would notice that there are many messages from Allah SWT about preserving and maintaining the environment yet, they do not give serious attention in taking care of their relationship with the environment instead, they are one of the reasons that contributed to the damage and destruction of nature.

Acknowledgment

This study is a part of fundamental research grant scheme (FRGS/1/2020/SS0/UMT/02/10) under MOHE (KPT) Malaysia conducted together with other researchers from UMT and UniSZA.

REFERENCES

[1] G. T. Miller, Jr, Living in the Environment – Principles, Connections and Solutions, 12th ed., USA: Brooks/ Cole – Wadsworth Thomson Learning, 2002.

[2] M. I. A. Al-Marbawi, Kamus Idris Al-Marbawi, Kuala Lumpur: Darul Nu'man, 1998, pp. 40.

[3] Quran (al-Isra') 17:70

[4] Quran (at-Tin) 95:4-6

[5] Quran (al-A'raf) 7:74

[6] Quran (an-Naml) 27:31

[7] Quran (al-A'raf) 7:85

[8] Quran (ar-Rum) 30:41

[9] A. J. Maidin, Religious and Ethical Values in Promoting Environmental Protection in the Land Use Planning System: Lessons for Asian Countries, In Issues in Islamic Law, by Abdul Haseeb Ansari, New Delhi, India: Serials Publications, 2007, pp.188-218.

[10] A. H. A. Rahman, S. Said, H. Salamun, H. Aziz, F. Adam, & W.I.W. Ahmad, Sustainable development from Islamic perspective, International Journal of Civil Engineering and Technology, vol. 9, 2018, pp. 985-992.

[11] M. Mohd Nor, S., M. Fatahiyah, & I. Ismaniza, Islamic Philosophy on Behaviour - Based Environmental Attitude, Journal of Social and Behavioral Sciences, vol. 49, 2012, pp. 85-92.

[12] N. Okour, Sustainable Development Environmental Values and in Islamic views, Journal of Economics and Sustainable Development, vol. 4, no. 14, 2013, pp. 136-145.

[13] H. Norizan, S. Hussin, & A. R. Hasimah, Sumbangan Teknologi Hijau Dalam Ketamadunan Islam, Sains Humanika, vol. 8, no. 3, 2016, pp. 29–37.

[14] D. Abdelzaher, A. Kotb, & A. Helfaya, Eco-Islam: Beyond the principles of why and what, and into the principles of how, Journal of Business Ethics, vol. 155, 2019, pp. 623–643.

[15] Y. Al-Qardawi, The Prophet and Science, Cairo: Rissala Foundation, 1995.

[16] Quran (az-Zumar) 39:21

[17] A. R. Haliza, Kesyumulan Islam dalam Aspek Pemeliharaan Alam Sekitar, Jurnal Sultan Alauddin Sulaiman Shah, 2019, pp. 353-365.

3D-Technologies in the Activities of the National Museum of Tajikistan

Mubirkhoni Muminpur

***Abstract*—** This article explores the National Museum of Tajikistan's experience with using information technology to create exhibitions. The goal of augmented reality is to improve the museum visitor's impression of the place by allowing them to participate in a specific activity. It provides a detailed analysis of several of the author's exemplary works on 3D scanning of museum objects from the author's collections for the exhibition "Heritage Corridors of the Silk Road in Afghanistan, Central Asia, and Iran – International Aspects of the European Year of Cultural Heritage" as part of a three-year EU-UNESCO project.

***Keywords*—** information technology, augmented reality, exposure, virtual exhibition, 3D.

Muminpur Mubirkhoni is with the National Museum of Tajikistan, Tajikistan (e-mail: mubeerkhon@gmail.com).

A Framework for Chinese Domain-Specific Distant Supervised Named Entity Recognition

Qin Long, Li Xiaoge

***Abstract*—**The Knowledge Graphs have now become a new form of knowledge representation. However, there is no consensus in regard to a plausible and definition of entities and relationships in the domain-specific knowledge graph. Further, in conjunction with several limitations and deficiencies, various domain-specific entities and relationships recognition approaches are far from perfect. Specifically, named entity recognition in Chinese domain is a critical task for the natural language process applications. However, a bottleneck problem with Chinese named entity recognition in new domains is the lack of annotated data. To address this challenge, A domain distant supervised named entity recognition framework is proposed. The framework is divided into two stages: first, the distant supervised corpus is generated based on the entity linking model of graph attention neural network; secondly, the generated corpus is trained as the input of the distant supervised named entity recognition model to train to obtain named entities. The link model is verified in the ccks2019 entity link corpus, and the F1 value is 2% higher than that of the benchmark method. The re-pre-trained BERT language model is added to the benchmark method, and the results show that it is more suitable for distant supervised named entity recognition tasks. Finally, it is applied in the computer field, and the results show that this framework can obtain domain named entities.

***Keywords*—**Distant named entity recognition, Entity linking, Knowledge graph, Graph attention neural network.

I. Introduction

The Knowledge Graphs depicts an integrated collection of real-world entities which are connected by semantically-interrelated relation. Therefore, KGs always to be used as the main means of tacking a plethora of real-life problems in various domains. However, there is no consensus in regard to a plausible and inclusive of a domain-specific entities and relationships. Named entity recognition is a basic task in natural language processing and attracts more and more attention. Although NER have achieved great process in the task of natural language processing, it still faces significant obstacles regarding domain named entity recognition. Domain named entity recognition can be described as the linguistic representations of domain specific concepts. NER is the task of detecting mentions of real-world entities from text and classifying them into predefined types. In recent years, more and more researches focus on NER in deep learning methods. With the development of deep learning, deep learning models have shown strong performance. However, most deep learning methods rely on large amounts of labeled training data. To tackle the label scarcity issue, the distant supervised named entity recognition is proposed. The labeling procedure is to match the tokens in the target corpus with concepts in knowledge bases in distant supervised named entity recognition. Nevertheless, distant supervised named entity recognition suffered from two major challenges: incomplete annotation and label noise issue. To solve above problems of distant supervision for NER, some researches have attempted to address it. Yang et al. [1] adopt the partial annotation CRFs to consider all possible labels for unlabeled tokens, but this method still require a considerable amount of annotated tokens or external tools; Cao et al. [2] attempt to induce labels for entity mentions based on their occurrence popularity in the concept taxonomy, which can suffer from labeling bias and produce mislabeled data; Liang et al. [3] studied the open-domain named entity recognition problem under distant supervision, which uses the pre-trained language model to improve the prediction performance of the named entity recognition model. With the large-scale pre-training language model rapidly becoming the mainstream method of natural language processing tasks [4]. The attention mechanism proposed by Vaswani et al. [5] establishes the position of BERT pre-training language model in natural language processing tasks [6].

Domain named entity recognition, also often referred to as Automatic term extraction (ATE), is the automated process of identifying terms in specialized texts. In recent years, it has become an important pre-processing step in many natural language processing tasks. Wang et al. [7] used two deep learning classifiers to extract terms; Amjadian et al. [8] proposed an efficient method of combining distributed representation with term extraction; Hatty et al. [9] constructed classification tasks by defining fine-grained terms and using basic neural networks; Kucza et al. [10] used recurrent neural network to term extraction by means of sequence labeling; Shah et al. [11] used unsupervised learning to extract terms in material science; Sajatovic et al. [12] proposed a topic modeling method to extract terms on individual documents; Kessler et al. [13] extracted architectural terms based on search engines and Wikipedia; Pollak et al. [14] proposed a method of extracting terms from literary corpus and automatically aligned terms to from a domain term dictionary; Terry et al. [15] proposed a new method of monolingual and multilingual labeling to generate dataset for downstream tasks; Terry et al. [16] once again introduced the search platform in the field of automatic term extraction, and made a detailed introduction to automatic

Qin Long, School of Computer Xi 'an University of Posts and Telecommunications, Xi' an, China. (e-mail:ql@stu.xupt.edu.cn).

Li Xiaoge, School of Computer Xi 'an University of Posts and Telecommunications, Xi' an, China. (e-mail:lixg@xupt.edu.cn).

term extraction; Kafando et al. [17] proposed an intelligent method for extracting biomedical terms from text documents and subsequent analysis, which combines statistical metrics and syntactic change rules to extract term variants from the corpus.

Noisy annotation is one of the major challenges in domain distant supervised named entity recognition. The same entity mention can be mapped to multiple entity types in the knowledge bases. For instance, the entity mention ‘Machine learning’ can be mapped to both ‘a book’ and ‘subject’ in the knowledge base. While existing methods can lead to many false-positive samples and hurt the performance of named entity recognition models. To solve noisy annotation in domain distant named entity recognition and a large scale of labeled corpus, we propose our model, a domain named entity recognition with distant supervision, which combine with the task of automatic term extraction and distant supervised named entity recognition. The method obtains labeled corpus by linking Wikipedia and domain paper entities, and applies the corpus to the distant supervised named entity recognition model.

In summary, we make the following contributions:

1)We propose a framework of domain distant supervised named entity recognition, and this framework is applied to obtain named entities in computer field.

2) We use graph attention neural network to solve the problem of entity link errors caused by multiple mappings of the same entity in the task of distantly supervised named entity recognition.

3)We demonstrate that re-pre-trained language model can also provide additional semantic information during the training process for distantly supervised named entity recognition.

II. Approach

Our approach is used to extract domain named entitles (see Figure 1). Specifically, we propose a two-stage training algorithm: In the first stage, Generation of distant labeled data based on graph attention neural network. In this stage, firstly, the entities of Wikipedia are extracted, and the entity information is used as the knowledge base. Then the entity graph is constructed based on the knowledge base. Secondly, the BERT language model is used as the input to extract the graph embedded representation of the entity nodes in the graph attention neural network, and then the text information and entity graph embedded representation information of the domain paper are used as features to carry out the entity linking task in the binary entity link model based on BERT. In the second stage, the results of entity linking are used as a small amount of labeled corpus, and the results of entity links are trained as entity dictionaries to generate labeled data by matching with a large number of unlabeled data as the input of distant supervised named entity recognition. We give more details on our approach in what follows.

A. Entity linking model based on graph attention neural network

(1) Entity extraction based on Wikipedia classification system

Wikipedia uses different classification methods for each classification object. Such as for the contents of general items are classified according to disciplines, for chronological accounts are classified by time, for character items use nationality and occupation. From the perspective of discipline knowledge classification system. The entries of Chinese Wikipedia are mainly divided into eight categories: Religion and belief、around the world、Humanities and Social Sciences、Life art and culture、Engineering technology and Applied Science、Nature and Natural Science、Chinese culture and Sociology. Each directory is subdivided into many secondary directories, and so on, until it is subdivided into specific items. Entries in the computer field can be obtained from the secondary directory computer science under engineering technology and applied science. Starting from computer science, the sub categories and subpages under this classification are recursively extracted to obtain the entity dictionary in the computer field as the named entity.

(2) Entity Graph

The entities in the computer field are extracted in Wikipedia, and the basic information of the entities is obtained, then the entity knowledge base is constructed by using the entity knowledge base.

The information of the entity is represented by the graph structure, and the graph embedded representation between the entity and its attributes is obtained through the graph attention network. This computer domain entity knowledge base is constructed by referring to the CCKS2019 entity link knowledge base, which include entity and its attribute value, subject-entity、subject-id、subject-type、alias、predicate and object value. The entity triples are defined as follows:

TABLE I
RELATIONSHIP DEFINITION OF ENTITY

Relationship type	tripe	example
alias	<entity,alias,value>	<GAN,alias,Generative Adversarial networks>
attribute	<entity,attribute,value >	<GAN,domain,deep learning>

The candidate entity node is taken as the central node, and the alias and attribute values are used as neighbor nodes to describe the knowledge of the central node. The entity graph of computer domain is constructed by using the defined triple. At the same time, in order to reduce the segmentation of the subgraph, all the description attributes of the candidate entities of each data in the knowledge base are spliced with the attribute value text, and the keywords are extracted by TF-IDF. If the extracted keywords coincide with the original attributes, only one is retained. The keywords are integrated into the entity graph(see Figure 2).

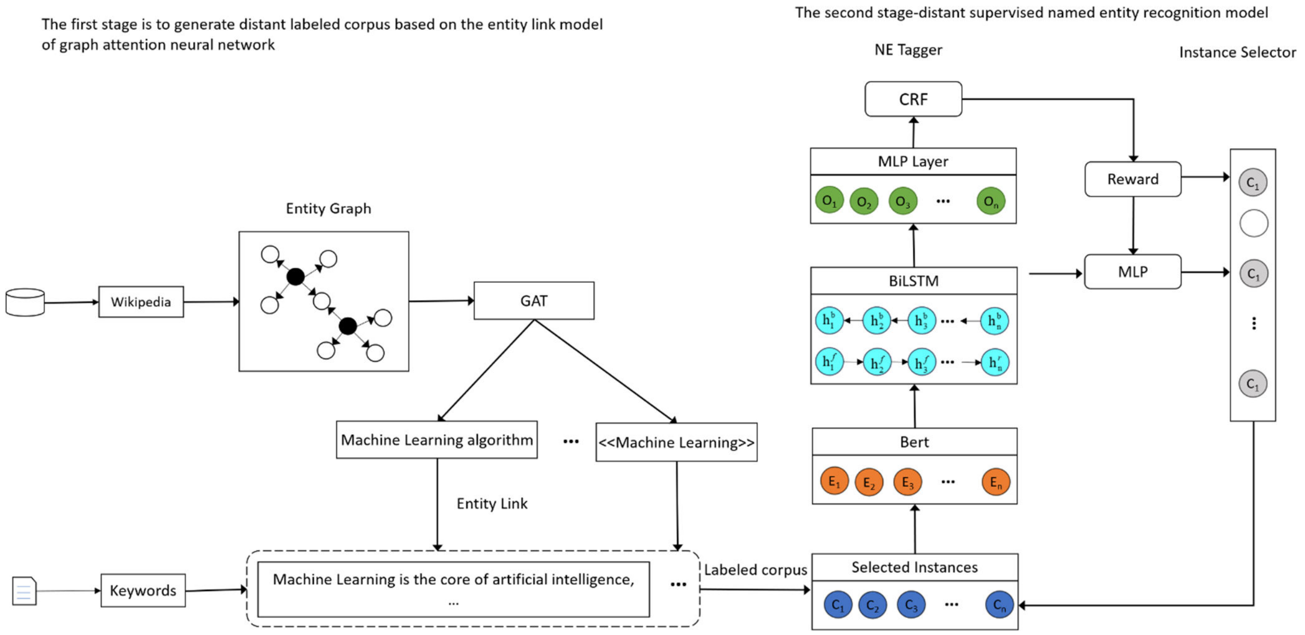

The first stage is to generate distant labeled corpus based on the entity link model of graph attention neural network
The second stage-distant supervised named entity recognition model
NE Tagger
Instance Selector
CRF
MLP Layer
Reward
MLP
Entity Graph
Wikipedia
GAT
BiLSTM
Machine Learning algorithm
<<Machine Learning>>
Bert
Entity Link
Keywords
Machine Learning is the core of artificial intelligence, ...
Labeled corpus
Selected Instances

Figure 1: The framework of domain distant supervised named entity recognition. We propose a two- stage training algorithm: In first stage, combine Wikipedia and domain paper to generate labeled data based on the entity link model of graph attention neural network; In the second stage, a distant supervised named entity recognition.

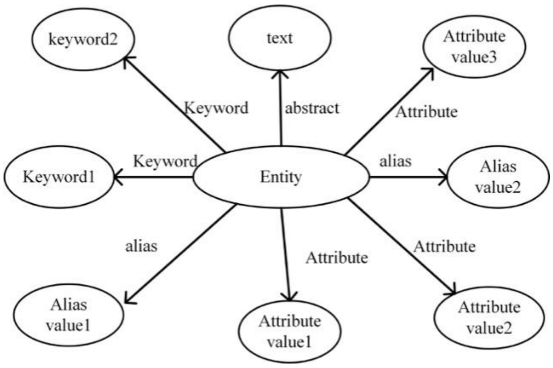


Figure 2: computer entity and its attribute

(3) Graph attention neural network

The data of graph structure contains two features: vertex feature and neighbor feature. GCN [18] cannot handle the problem of dynamic graphs, and it is not easy to assign different weight to different neighbors. However, GAT consider the structure of the graph in space, so it can perfectly adapt to the dynamic graph. Velickovic et al. [19] proposed graph attention neural network, and many researchers have applied it to a variety of natural language processing tasks and achieved good results; Zhang et al. [20] build graph information for documents and obtain fine-grained word representations of their local structures for text classification; Huang et al. [21] used graph neural network for text classification; Zheng et al. [22] proposed a new multi-granularity machine reading comprehension framework in machine reading comprehension, which uses graph attention network to obtain different levels of representation so that they can be learned at the same time.

The entity graph is constructed according to the triple relationship of entities and their attributes defined in Table 1 and the keywords added in figure 2. The entity graph is studied by embedding the graph into the node representation through the graph attention neural network. Using the attention mechanism, the characteristics of neighboring nodes are weighted and summed, and different weights are assigned to different neighboring nodes according to their characteristics. The graph attention neural network model is as Figure 3.

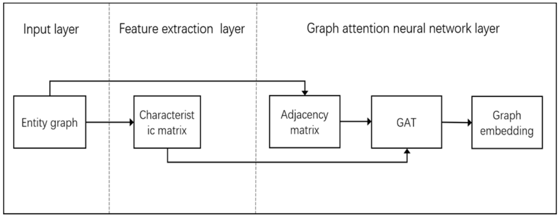


Figure 3: Graph attention mechanism Neural Network and its graph embedding.

The model consists of input layer, feature extraction layer and graph attention neural network layer.

The input layer is entity graph, which consists of a set of X nodes $V=\{V_1,V_2...V_X\}$ and Y edges $E=\{E_1,E_2...E_Y\}$. In the feature extraction layer, the BERT pre-training model is used to extract the feature of the node text in the entity graph, and the $m\times n$ feature matrix is obtained, where m is the number of graph nodes and n is the feature dimension. The graph attention layer obtains the feature matrix and the adjacency matrix from the input layer and the feature extraction layer, and inputs them to the two-layer graph attention neural network for representation. The final output is the graph node feature representation represented by GAT.

The following is a detailed description of the GAT model.

The input to GAT is a set of node features, $h=\left\{\vec{h}_1,\vec{h}_2...\vec{h}_N\right\}$, $h_i \in \mathbb{R}^F$, where N is the number of nodes, and F is the number of features in each node. $h'=\left\{\vec{h}'_1,\vec{h}'_2...\vec{h}'_N\right\}$, $\vec{h}_i' \in \mathbb{R}^F$, as its output.

The coefficients computed by the attention mechanism may be expressed as:

$$\alpha_{ij} = softmax\left(\sigma\left(\vec{a}^T\left[W\vec{h}_i \| W\vec{h}_j\right]\right)\right) \tag{1}$$

Where α_{ij} represents the attention coefficient between nodes i and j , $W \in \mathbb{R}^{F\times F'}$, is a weight matrix, $\vec{h}_i$ and $\vec{h}_j$ represents the node characteristics of node i and j, the dimension of $\vec{h}_i$ is $1\times F$, then the dimension of $w\vec{h}_i$ is $1\times F'$, $\|$ represents concatenation, and the tensors of two $1\times F'$ are glued together to a large tensor of$1\times 2F'$. Then by multiplying the dimension of the attention convolution kernel coefficient $\vec{\alpha}^T$, and the dimension of $\vec{\alpha}^T$ is $2F'\times 1$,the final result is a number. σ indicates the activation function, which is LeakyReLU. After activating the function, the final attention result is calculated by softmax.

$\vec{h}_i$ comes from the following formula:

$$\vec{h}_i' = \sigma\left(\sum_{j\in \mathrm{N}_i} \alpha_{ij} W\vec{h}_j\right) \tag{2}$$

$\vec{h}_i$ represents the output characteristics of the layer with respect to node i, where N_i is some neighborhood of node i in the graph. Specifically, multi-head attention to be beneficial for mechanism. Where k independent attention mechanisms execute the transformation of Equation 2, and then their features are concatenated, so the output feature representation:

$$\vec{h}_i' = \|_{k=1}^{\mathrm{K}} \sigma\left(\sum_{j\in \mathrm{N}_i} \alpha_{ij} W\vec{h}_j\right) \tag{3}$$

The final output layer can be represented as:

$$\vec{h}_i' = \sigma\left(\frac{1}{\mathrm{K}}\sum_{k=1}^{\mathrm{K}}\sum_{j\in \mathrm{N}_i} \alpha_{ij}{}^k W^k \vec{h}_j\right) \tag{4}$$

(4) Entity link

A large number of researchers have improved the BERT model in the task of entity linking. Zhan et al. [23] proposed to introduce the BERT pre-training language model into the entity link task to analyze the context of the entity reference item and the relevant information of the candidate entity. By improving the effect of semantic analysis to enhance the results of entity links, and using TextRank keyword extraction technology to enhance the topic information of target entity comprehensive description information, and enhance the accuracy of text similarity measurement; Luo et al. [24] combine global and local attention mechanism to obtain the hidden state of the text.

In 2019, the National Conference on knowledge Graph and semantics issued the task of entity linking evaluation for Chinese short texts. The best solution uses dictionary matching to get the entities in the short text, and finally uses the BERT-EntityNameEmbedding (BERT-ENE) [1] model to filter the results, so as to achieve entity recognition. In the part of entity link, the two-classification model based on BERT is used to predict and rank the candidate entities.

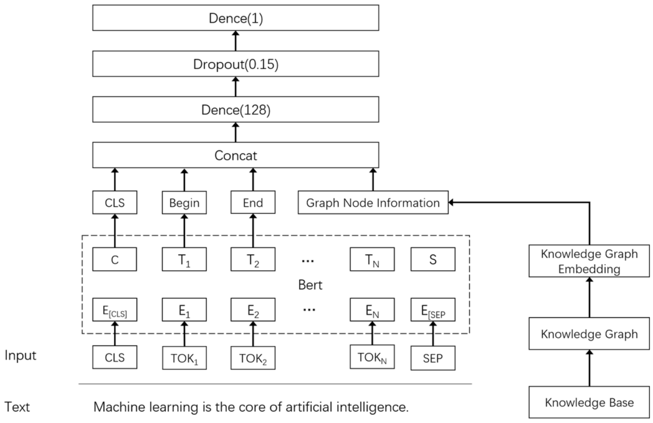


Figure 4: Entity link model based on BERT and Graph embedding.

The entity linking model of our framework is shown in figure 4, using binary entity links based on BERT. We consider node feature representation of entities in graph attention mechanism network. Short text and description text of the entity to be lined as input of model. The features are the BERT vector of the text, the start and end position vectors of the candidate entity, and the node feature representation of the entity in the graph attention mechanism neural network. The four feature vectors are spliced, and after the full connection layer, the probability scores of the candidate entities are sorted by sigmoid activation, and the correct entity with the highest probability is selected.

B. Named entity recognition with distant supervised

The Chinese language pre-training model has strong performance in natural language processing tasks. Gururangan et al. [25] proposed that language model pre-training can greatly improve the effect of subtasks. For example, continuing pre-training on the data set of specific tasks can improve the effect very cheaply. The effect can be improved by continuing pre-training on the data set of the target domain, and the more irrelevant the corpus of the target domain and the original training corpus is, the more obvious the improvement effect is.

A number of pre-training models have also been produced in the field of Chinese pre-training models, such as the Chinese BERT model[2], the Roberta pre-training model released by IFLYTEK of Harbin University of Technology[3], and the ERNIE pre-training model released by Baidu[4].

In order to study the performance of the pre-training model in the task of distantly supervising named entity recognition, on the basis of three Chinese pre-training models, People's Daily corpus, Amazon commodity review corpus and restaurant review corpus are used for retraining.

[1] https://github.com/panchunguang/sskc_baidu_entity_link
[2] https://huggingface.co/bert-base-chinese/tree/main
[3] https://github.com/ymcui/Chinese-BERT-wwm
[4] https://github.com/nghuyong/ERNIE-Pytorch

After obtaining the computer domain entity, the labeled corpus is obtained by matching the entity and unlabeled corpus, which is applied to the distant supervision of named entity recognition task (see figure 1).

Following Yang et al. (2018)[5], we consider BERT pre-training language model on the distant supervised entity recognition model. The following is a detailed introduction of this model.

As shown in Figure 5, the model of distant supervised named entity recognition consists of two modules: the NE Tagger built on the idea of partial annotation learning to reduce the effect of unknown-type characters, the instance selector which choose positive sentences from a large scale of unlabeled corpus and provides them to the NE tagger to train model.

Initially, we get a small set of labeled data D from the entity linking model and a large scale of unlabeled data U. According to the results of entity linking, we collect named entity to construct dictionary E about computer field, then using the entries of E to match the sentences in U by the method of distant supervision. And we also obtain a set of sentences containing at least matched string, and the set is called I. According to the traditional BIO schema to represent the tags of sentences, the beginning character of an entity in I are marked with "B-XX", "I-XX" is used to mark other characters of the entity, and the character as "O" if it is not in the entity.

LSTM-CRF-PA

It is a common problem known as false negative instance in distant supervised named entity recognition. If we arbitrarily label as "O" which may misguide model to learn the false instance. Therefore, each non-matched character should be tagged as the appropriate label. A set of label sequences z for every distantly supervised sentence, whose probability is naturally the sum of probability of each possible label sequence $\tilde{y}$ in z. Therefore, the probability of the distantly supervised instance is calculated as:

$$p(z \mid x)=\sum_{\tilde{y} \in z} p(\tilde{y} \mid x)=\frac{\sum_{\tilde{y} \in z} e^{score(x, \tilde{y})}}{\sum_{\tilde{y} \in Yx}^{n} e^{score(x, \tilde{y})}} \tag{5}$$

The loss function of the model with CRF-PA cam be computed as follows:

$$loss(\theta, x, z)=-\log p(z \mid x) \tag{6}$$

Instance Selector for Noisy Annotation

To train an agent as an instance selector with reinforcement learning technology. The initial labeled data D and the distantly supervised data I is denoted as a candidate dataset A. At each episode, we collect a random-size package of instances B from A. By default, all the supervised instances in the current package are selected without decisions of agent. For each distantly supervised instances in the current package, the agent performs an action from the {1,0} to decide whether to select this instance. The agent will be rewarded when all actions are completed. The reward represents action feedback on this package and will be used to update the agent.

State representation

The vector s_{t} represents the current instance and its label sequence, which consists of two information: first, the vector representation of the output from BiLSTM layer. Secondly, the label score calculated with output of the MLP layer from the shared encoder and annotation of this instance.

Policy network

The agent determines whether the behavior selector will select the $t-th$ distantly supervised instance. Then we use a logistic function as the policy function:

$$A_{\theta}(s_t, a_t)=a_t \sigma(W * S_t+\mathrm{b})+(1-\mathrm{a}_{\mathrm{t}})(1-\sigma(W * S_t+\mathrm{b})) \tag{7}$$

Reward

The reward is used to evaluate the ability of current NE tagger to predict labels of each character. The model receives a delayed average reward when it completes all elections in current package, and before that the reward for each action is zero. The current package A consists of two subsets: The labeled data D from entity linking and B from the distantly supervised instances. The NE tagger calculates the probability of each sentence in A. The reward can be calculated on selected distantly supervised instances $\tilde{A}$ and the labeled data:

$$r=\frac{1}{|\tilde{A}_s|+|\tilde{H}|}\left(\sum_{x_{j, z \in \tilde{A}_S}} \log p(z \mid x_j)+\sum_{x_{j, z \in \tilde{H}}} \log p(y \mid x_k)\right)$$

(8)

Selector training

In order to maximize the reward of the selections, we use policy gradient method to optimize the policy network. For each random-size package A, the feedback for each action is the same as the average reward r. Then we calculate the gradient and update the selector.

$$\theta=\theta+\alpha \sum_{t=1}^{|\tilde{A}|} r(a_t) \nabla_{\theta} \log A_{\theta}(s_t, a_t) \tag{9}$$

III. Experimental results

A. Dataset

Entity link. We use the evaluation data set provided by the CCKS2019 entity link task for Chinese short text as the verification data of the entity link model. Its dataset includes 90000 pieces of labeled data and 39925 pieces of knowledge base information.

Distant supervised named entity. Following Yang et al. (2018), the news corpus dataset is selected, 3000 sentences were randomly selected as the training corpus, 3328 as the verification corpus and 3186 as the test corpus. In order to verify the effect of corpus retraining on BERT pre-training model, People's Daily 2016 data set, Amazon product review data and food review data were used as pre-training corpus and retrained on three pre-training models: BERT, Roberta and ERNIE.

Computer domain entity linking and distant supervised named entity recognition. Wikipedia is used to build a computer domain entity knowledge base, which contains a total

[5] https://github.com/rainarch/DSNER

of 84493 pieces of data. The entity link data is the abstract of the paper, and 201608 pieces of data are extracted.

After linking through the entity, the named entity is matched in the untagged corpus to get 9000 pieces of data, which are divided into training, verification and testing remote supervised named entity recognition model according to 1:1:1.

B. Entity linking experiments

In the entity linking experiment, the entity linking task of ccks2019 is used to verify. On the basis of the first prize in the ccks2019 contest, the BERT-ENE model of the original author is adopted in the entity recognition module. In the entity linking module, the knowledge base information is transformed into graph information, and the link model is realized by using short text and entity embedding based on graph attention mechanism. The results are shown in the following table II.

TABLE II
RESULT OF ENTITY LINK (%)

model	F1
EBRT	80.13
BERT-Graph	83.29

As can be seen from Table II, the effect of adding entity graph embedding model is better than that of text input based on BERT. It shows that through graph attention neural network, we can obtain more semantic information about the entity, its attributes and its text.

C. Distant named entity recognition experiments

A distant supervised named entity recognition model based on BERT pre-training language model is adopted. In order to compare with the word2vec of the benchmark model, the first 100-dimensional vector of the output of the first layer of BERT is selected as the feature. And under the People's Daily corpus, Amazon commodity review corpus and food review corpus, the Chinese BERT model, Roberta model and ERNIE model are pre-trained again. The retraining corpus training set uses data about the size of 300MB.

TABLE III
RESULT OF WORD2VEC AND CHINESE BERT

model	DEV			DEV		
	P	R	F1	P	R	F1
word2vec-100	86.94	80.12	83.40	81.63	76.95	79.22
BERT-100	89.64	80.73	84.95	85.48	77.39	81.23

As shown in Table III, the performance of the BERT pre-training language model is obviously better than that of the word2vec model. The effect of retraining the three Chinese pre-training language models on the People's Daily corpus is shown in Table IV.

TABLE IV
RESULTS OF RE-PRE-TRAINING MODEL ON THE PEOPLE'S DAILY CORPUS

model	DEV			DEV		
	P	R	F1	P	R	F1
word2vec-100	86.94	80.12	83.40	81.63	76.95	79.22
BERT-100	87.92	81.76	84.73	84.04	79.18	81.54
Roberta-100	86.14	82.20	84.12	81.34	79.88	80.60
ERNIE-100	87.00	80.47	83.61	83.73	76.39	79.90

Among the three Chinese pre-training models under People's Daily corpus, Chinese BERT model is better than other models. The retraining effects of the three Chinese pre-training language models on Amazon commodity corpus and restaurant review corpus are shown in Table V and Table VI respectively.

TABLE V
RESULTS OF RE-PRE-TRAINING MODEL ON THE AMAZON COMMODITY CORPUS

model	DEV			DEV		
	P	R	F1	P	R	F1
word2vec-100	86.94	80.12	83.40	81.63	76.95	79.22
BERT-100	86.98	82.02	84.43	84.19	79.58	81.82
Roberta-100	81.30	81.15	81.23	80.04	77.09	78.54
ERNIE-100	83.17	81.59	82.37	79.98	77.59	78.77

TABLE VI
RESULTS OF RE-PRE-TRAINING MODEL ON THE RESTAURANT REVIEW CORPUS

model	DEV			DEV		
	P	R	F1	P	R	F1
word2vec-100	86.94	80.12	83.40	81.63	76.95	79.22
BERT-100	87.25	82.80	84.97	83.68	80.18	81.89
Roberta-100	84.02	81.33	82.65	80.35	78.59	79.46
ERNIE-100	85.73	80.47	83.01	82.07	75.70	78.76

After retraining under the commodity review corpus and restaurant reviews, the effect of Chinese BERT pre-training was significantly improved, but the sub-accuracy of Roberta and ERNIE models decreased significantly, indicating that a lot of semantic information was lost after retraining.

In order to verify the influence of different size training corpus on the pre-training language model, the experimental results of increasing the training corpus are shown in Table VII.

TABLE VII
RESULTS OF RE-PRE-TRAINING MODEL ON THE DIFFERENT SIZES OF RESTAURANT REVIEW CORPUS

corpus	DEV			DEV		
	P	R	F1	P	R	F1
306MB	87.25	82.80	84.97	83.68	80.18	81.89
616MB	88.83	82.45	85.82	83.97	79.28	81.56
919MB	87.71	82.02	84.77	84.10	78.49	81.20
1.3GB	88.47	81.59	84.89	83.93	76.99	80.31

As shown in Table VII, increasing the training corpus can significantly improve the performance of remote supervised named entity recognition, but with the increase of data, the effect tends to be stable.

D. Computer domain experiments

In order to verify the applicability of the model proposed in our paper, we use Wikipedia and computer paper data for entity acquisition in the computer domain. Entity linking adopts two-classification model based on BERT+ graph embedding.

TABLE VIII
RESULTS OF COMPUTER DOMAIN ENTITY LINK

model	F1
BERT-Graph	90.34

The results of computer remote supervision of named entity recognition are shown in Table IX.

TABLE IX
RESULTS OF COMPUTER DOMAIN

model	DEV			DEV		
	P	R	F1	P	R	F1
BERT	92.14	90.56	91.34	90.33	89.55	89.94
BERT-sp	95.36	90.78	93.01	93.57	90.23	91.86

As can be seen from Table IX, the BERT model uses Chinese BERT as input, and BERT-sp is the result of re-training using commodity review 616MB training corpus on the basis of Chinese BERT pre-training. The results show that the effect of re-pre-training is increased in all indicators. Therefore, the effect of re-pre-training can improve the performance of the task in the distant supervision of named entity recognition task.

IV. CONCLUSION

Recent work in natural language processing has focused on domain named entity recognition. In this paper, we show a distant supervised named entity recognition, which combines Wikipedia and paper data to obtain labeled corpus, and applies the labeled data to distant supervised named entity recognition. Finally, the method is applied to the computer domain entity recognition, and the experimental results show that the method can adapt to the domain named entity recognition task.

We use the entity graph and the paper to carry on the entity link, and the entity link model is represented by the embedded representation of the text and the entity graph. As semi-structured data, the paper data can also be used to build a graph, and the next step is to link the entity graph with the paper graph. Secondly, in the graph attention neural network, adding PMI to the graph pruning to obtain the important attributes of the entity can obtain a better entity graph embedding representation to improve the performance of the entity link.

REFERENCES

[1] Yang, Y., Chen, W., Li, Z., and Zhang, M, ‘Distanly supervised NER with partial annotation learning and reinforcement learning,’ Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, New Mexico, USA, August 2-26, pp. 2159-2169, 2018.

[2] Cao, Y., Hu, Z., Chua, T. S., Liu, Z., and Ji, H, ‘Low-resource name tagging learned with weakly labeled data,’ arXiv preprint arXiv:1908.09659.

[3] Liang, C., Yu, Y., Jiang, H., Er, S., Wang, R., Zhao, T., and Zhang, C, ‘Bond: Bert-assisted open-domain named entity recognition with distant supervision,’ In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Virtual Event, USA, August 23-27, pp. 1054-1064, 2020.

[4] Devlin, J., Chang, M. W., Lee, K., and Toutanova, K, ‘BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,’ Proceedings of NAACL-HLT, pp. 4171-4186, 2019.

[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., and Polosukhin, I. ‘Attention is all you need,’ Advances in neural information processing systems, pp. 5998-6008, 2017.

[6] Peters M E, Neumann M, Iyyer M. ‘Deep contextualized word representations,’ Proceedings of NAACL-HLT, pp. 2227-2237.

[7] Wang, R., Liu, W., and McDonald, C, ‘Featureless domain-specific term extraction with minimal labelled data,’ Proceedings of the Australasian Language Technology Association Workshop 2016, pp. 103-112, 2016.

[8] Amjadian, E., Inkpen, D., Paribakht, T., and Faez, F, ‘Local-global vectors to improve unigram terminology extraction,’ Proceedings of the 5th International Workshop on Computational Terminology, pp. 2-11, 2016.

[9] Hätty, A., and im Walde, S. S, ‘Fine-grained termhood prediction for german compound terms using neural networks,’ Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions, Santa Fe, New Mexico, USA, August 25-26, pp. 62-73, 2018.

[10] Kucza, M., Niehues, J., Zenkel, T., Waibel, A., and Stüker, S, ‘Term Extraction via Neural Sequence Labeling a Comparative Evaluation of Strategies Using Recurrent Neural Networks,’ In Interspeech, pp. 2072-2076, 2018.

[11] Shah, S., and Reddy, S, ‘Similarity Driven Unsupervised Learning for Materials Science Terminology Extraction,’ Computación y Sistemas, vol. 23. pp. 1005-1013, 2019.

[12] Šajatović, A., Buljan, M., Šnajder, J., and Bašić, B. D. ‘Evaluating automatic term extraction methods on individual documents,’ Proceedings of the Joint Workshop on Multiword Expressions and WordNet, pp. 149-154, 2019.

[13] Kessler, R., Béchet, N., and Berio, G, ‘Extraction of terminology in the field of construction,’ 2019 First International Conference on Digital Data Processing, pp 22-26, 2019.

[14] Pollak, S., Repar, A., Martinc, M., and Podpečan, V, ‘Karst exploration: extracting terms and definitions from karst domain corpus,’ Proceedings of eLex, pp. 934-956, 2019.

[15] Terryn, A. R., Hoste, V., and Lefever, E, ‘In no uncertain terms: a dataset for monolingual and multilingual automatic term extraction from comparable corpora,’ Language Resources and Evaluation, vol.54, pp.384-418, 2020.

[16] Rigouts Terryn, A., Hoste, V., Drouin, P., and Lefever, E, ‘Termeval 2020: Shared task on automatic term extraction using the annotated corpora for term extraction research (acter) dataset,’ 6th International Workshop on Computational Terminology, pp. 85-94, 2020.

[17] Kafando, R., Decoupes, R., Valentin, S., Sautot, L., Teisseire, M., and Roche, M, ‘ITEXT-BIO: Intelligent Term EXTraction for BIOmedical Analysis,’ Health Information Science and Systems, vol. 9, pp. 1-23, 221.

[18] Kipf, T. N., and Welling, M, ‘Semi-supervised classification with graph convolutional networks,’ arXiv preprint arXiv:1609.02907, 2016.

[19] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y, ‘Graph attention networks,’ arXiv preprint arXiv:1710.10903, 2017. *arXiv preprint arXiv:1710.10903.*

[20] Zhang, Y., Yu, X., Cui, Z., Wu, S., Wen, Z., and Wang, L, ‘Every Document Owns Its Structure: Inductive Text Classification via Graph Neural Networks,’ Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 334-339, 2020. *arXiv preprint arXiv:2004.13826.*

[21] Huang, L., Ma, D., Li, S., Zhang, X., and Wang, H, ‘Text Level Graph Neural Network for Text Classification,’ Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 3444-3450, 2019. *arXiv preprint arXiv:1910.02356.*

[22] Zheng, B., Wen, H., Liang, Y., Duan, N., Che, W., Jiang, D., and Liu, T, ‘Document Modeling with Graph Attention Networks for Multi-grained Machine Reading Comprehension,’ Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 678-6718. *arXiv preprint arXiv:2005.05806.*

[23] Zhan fei, ZHU Yanhui, and LIANG Wentong, ‘Entity Linking Via BERT and TextRank Keyword Extraction,’ Journal of Hunan University of Technology,vol.34, pp. 63-70, 2020.

[24] Luong, M. T., Pham, H., and Manning, C. D, ‘Effective approaches to attention-based neural machine translation,’ ArXiv Preprint ArXiv:1508.04025, 2015.

[25] Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., and Smith, N. A, ‘Don't stop pretraining: adapt language models to domains and tasks,’ arXiv preprint arXiv:2004.10964, 2020.

BERT- Based Chinese Coreference Resolution

First A. Li Xiaoge, Second B. Wang Chaodong.

Abstract**—**We introduce the first Chinese Coreference Resolution Model based on BERT (CCRM-BERT) and show that it significantly outperforms all previous work. The key idea is to consider the features of the mention, such as part of speech, width of spans, distance between spans, etc. And the influence of each features on the model is analyzed. The model computes mention embeddings that combine BERT with features. Compared to the existing state-of-the-art span-ranking approach, our model significantly improves accuracy on the Chinese OntoNotes benchmark.

Keywords**—** BERT, coreference resolution, deep learning, nature language processing

I. Introduction

Coreference resolution refers to the problem of determining with noun phrases (NPs) refer to each real-world entity mentioned in a document. We present the first Chinese neural coreference resolution model that adds the features to spans representations. All recent coreference models, including end-to-end neural coreference resolution [1] that achieved impressive performance gains, ignore features of the mention. We demonstrate for the first time that these resources are required, and in fact performance can be improved significantly with them. The features are crucial signals for coreference resolution.

At the core of our model are vector embeddings representing spans of text in the document, which combine BERT with features of spans. BERT-based models have reported dramatic gains on multiple semantic benchmarks including question-answering, natural language inference, and named entity recognition [2]. The model with features can make better decisions, for each span, which of the previous spans (if any) is a good antecedent. In our analyses, we experimentally that these features correlate strongly with the accuracy of our model.

Scoring all span pairs in our model is impractical, therefore we factor the model over unary mention scores and pairwise antecedent scores, both of which are simple functions of the learned span embedding [1].

We propose a BERT-based Chinese coreference resolution model for the Chinese coreference resolution problem, and add the features of spans to spans representations. Our final approach outperforms existing models by 1.1 F1 on the Chinese OntoNotes benchmark.

II. Related Work

Early reference resolution model was rule-based, of which two are the most representative. Hobbs algorithm [3] proposed by Hobbs is the earliest reference resolution method, which is a method of pronoun resolution based on grammar analysis tree. The central theory algorithm [4] is a partial discourse coherence theory proposed by Grosz et al., whose main idea is to locate the entity "focus" in the text. Up to now, many scholars have proposed different solutions.

Scoring span or mention pairs has perhaps been one of the most dominant paradigms in coreference resolution. The base coreference model used in this paper from Lee et al. [1] belongs to this family of models. A year later, Lee et al. [5] expanded on his work in two ways, first allowing the elaboration of all candidate word representations through iterative and gated attention mechanisms, and second using modifiable scoring mechanisms for pruning potential antecedents.

Fei et al. in 2019 proposed an incremental reference resolution model [6] that allows end-to-end training of the model using a supervised reference resolution task and an auxiliary language modeling task by encoding storage operations as distinguishable gating.

Joshi et al. in 2019 first applied BERT to a reference resolution model[2] replacing the full coding layer using BERT based on the higher-order reference resolution model of Lee et al[5]. and analyzing the difference in the role of different language models on reference resolution.

For the problem of reference resolution in Chinese, Xi et al [7] proposed a Deeping Learning mechanism for DBN models, using a multilayer unsupervised RBM network and a layer of supervised BP network to obtain text semantic features for reference resolution. Lv et al [8] proposed a reference resolution method based on the semantics of Chinese frames. The chapters annotated under the Chinese Frame Net (CFN) resource were preprocessed with data, after which the representations were subjected to reference resolution using a classification model.

III. Task

We formulate the coreference resolution task as a set of antecedent assignments y_i for each of span i in the given document, following Lee et al[1]. The set of possible assignments for each y_i is $y_i = \{\varepsilon, 1, \ldots, i-1\}$,a dummy antecedent ε and all preceding spans. The dummy antecedent ε represents two possible scenarios: (1) the span is not an entity mention or (2) the span is an entity mention but it is not coreferent with any previous span. These decisions implicitly define a final clustering, which can be recovered by grouping together all spans that are connected by the set of antecedent predictions [1].

IV. Model

For our model, we use BERT in combination with features

Chaodong Wang is with the Xi 'an University of Posts and Telecommunications, China (e-mail: 941020637@qq.com).

of spans to represent word embeddings. The model architecture is shown in Figure 1, and we will present it layer by layer below.

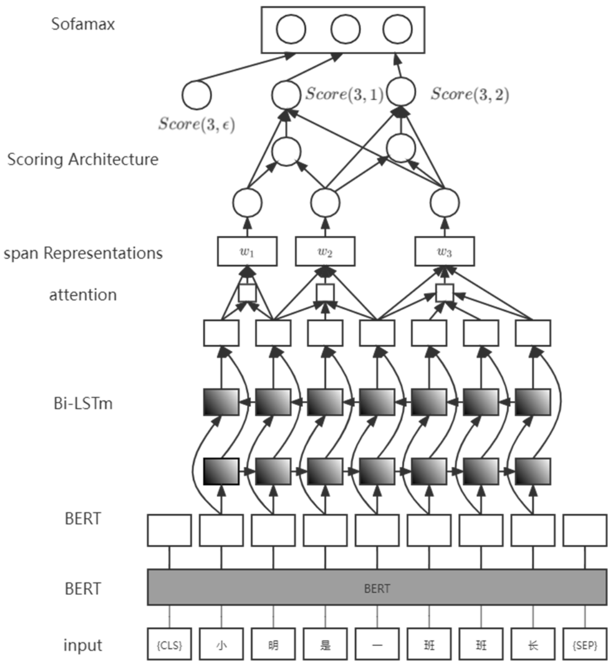


Figure 1: First step of our model, which computes embedding representations of spans for scoring potential entity mentions. Second step of our model, which computes the mention scores. The final coreference score of a pair of spans is computed by summing the mention scores of both spans and their pairwise antecedent score.

A. BERT

The purpose of the BERT layer is to obtain context-dependent vectors and to make the resulting high-dimensional vector representation available for subsequent functional components.

There are two traditional representations for obtaining text vectors: one-hot encoding and the classical language model Word2Vec [9] based on contextual information, but since the word vectors are determined once the training is completed, they cannot effectively distinguish between cases of multiple meanings of a word. The BERT based on bidirectional Transformer [10] can generate the corresponding contextualized vector representation based on the contextual information to solve the problem of multiple meanings of words. Apart from better bidirectional reasoning, one of BERT's major improvements over previous methods is passage-level training, which allows it to better model longer sequences [11].

B. Bi-LST

As an improved version of recurrent neural networks, LSTM mainly solves the gradient disappearance problem that occurs in RNNs with longer utterances. Since both RNN and LSTM are one-way temporal, only semantic information of past moments can be obtained. In order to make the characters obtain their antecedent and posterior semantic information at any moment, this model chooses to use a bidirectional LSTM neural network[12]:

$$i_{t,d} = \sigma(W_i[x_t, h_{t+d,d}] + b_i) \quad (1)$$

$$f_{t,d} = \sigma(W_f[x_t, h_{t+d,d}] + b_f) \quad (2)$$

$$o_{t,d} = \sigma(W_o[x_t, h_{t+d,d}] + b_o) \quad (3)$$

$$\tilde{C}_{t,d} = \tanh(W_C[x_t, h_{t+d,d}] + b_C) \quad (4)$$

$$C_{t,d} = i_{t,d} \circ \tilde{C}_{t,d} + f_{t,d} \circ C_{t+d,d} \quad (5)$$

$$h_{t,d} = o_{t,d} \circ \tanh(C_{t,d}) \quad (6)$$

$$x_t^* = [h_{t,-1}, h_{t,1}] \quad (7)$$

Where indicates the positive and negative directions of the LSTM module and is the Sigmoid activation function. At moment , is the output of the corresponding input gate, is the output of the forget gate, is the vector of candidates, is the output after updating the cell status, is the output of the LSTM, , , , is the weight calculated for each variable. Finally, the output in both forward and reverse directions is spliced to generate as the final output of the LSTM layer.

C. Attention

This layer solves the problem of weighting different words by introducing an attention mechanism. Define w_i as the i-th candidate word in text T, and $first(i)$ and $last(i)$ as the indexes of the start and end in. Feed Forward Neural Network (FFNN) calculates the weight of each character and weights the weights of each character to obtain the attention vector $\tilde{w}_i$ representation of the candidate word w_i [15]:

$$\alpha_t = W_\alpha \cdot \text{FFNN}_\alpha(x_t^*) \quad (8)$$

$$\alpha_{i,j} = \frac{e^{\alpha_j}}{\sum_{k=first(i)}^{last(i)} e^{\alpha_k}} \quad (9)$$

$$\tilde{w}_i = \sum_{j=first(i)}^{last(i)} \alpha_{i,j} \cdot x_j \quad (10)$$

where $\delta(i)$ is the width feature and part of speech feature of the candidate $\tilde{w}_i$ (these features will be described in detail later).

D. Span Representations

Combining the word representations x_t^* of the above content mention with the attention vector $\tilde{W}_i$, the final representations of the candidate words are obtained as follows:

$$m_i = \left[\tilde{w}_i, x^*_{first(i)}, x^*_{last(i)}, \delta(i)\right] \quad (11)$$

where $\delta(i)$ is the width feature and part of speech feature of the candidate $\tilde{w}_i$ (these features will be described in detail later).

E. Scoring Architecture

Given these span representations, the scoring functions above are computed by FFNN:

$$Score_m(i) = W_m \cdot \mathrm{FFNN}_m(m_i) \quad (12)$$

where □ denotes the dot product, and FFNN denotes a feed-forward neural network that computes a nonlinear mapping from input to output vectors.

After getting all possible mentions, it is necessary to determine two by two whether they are a pair of possible mention pairs. Let there exist an ordered set $\omega = \{w_1, w_2, ..., w_k\}$ of k candidate words for text , where if for w_i , w_j there exists $i > j$, then there must be $frist(i) > first(j)$,and $last(i) > last(j)$ if $first(i) = first(j)$. For all possible (w_i, w_j) , the mention pairs representations $p_{i,j}$ are calculated as follows:

$$p_{i,j} = [m_i, m_j, m_i \circ m_j, \delta(i,j)] \quad (13)$$

where $\circ$ is the Hadamard product, and the feature represented by $\delta(i,j)$ will be described in detail later. Finally, all possible mention pair scores are calculated by FFNN:

$$Score_p(i,j) = W_p \cdot \mathrm{FFNN}_p(p_{i,j}) \quad (14)$$

The task goal of obtaining all co-referential chains in the text is equivalent to obtaining the most likely antecedent word $y_i \in Y_i = \{\varepsilon, w_1, w_2, ..., w_{i-1}\}$ for each candidate word w_i in ω, where ε is an artificially added special element. If $y_i = \varepsilon$, it means that w_i has the following possibilities: w_i is not a mention; w_i is a mention but no antecedent.

For all the candidate words w_i to be confirmed, the co-reference scores need to be calculated one by one with the elements in the corresponding set of possible antecedents Y_i to obtain the most likely antecedent y_i for w_i . Let (w_i, w_j) be the candidate word pair to be calculated, and three sub-problems need to be considered: whether w_i is a mention; whether w_i is a mention; and whether w_j is the anteceded of w_i . Combining the mention score and the mention pairs score, the co-referent score is calculated as follows:

$$Score_{cr}(i,j) = \begin{cases} 0 & w_j = \varepsilon \\ Score_m(i) + Score_m(j) + Score_p(i,j) & w_j \neq \varepsilon \end{cases} \quad (15)$$

F. Features

We add different features to the mention pairs representations and span representations respectively. In span representations, the mention width and the part of speech were added, this is because considering a reasonable mention that is not too long and usually ends with a noun, pronoun or noun phrase, this has a positive effect on determining whether the candidate is a mention or not. For part of speech, mention may have multiple part of speech combinations, so when representing mention features, we refer to the idea of one-hot coding: if the mention appears in a certain part of speech, it is set to 1 in the corresponding part of speech flag position and 0 in the non-appearing part of speech flag position. Considering that the original data distribution is not changed, the z-core is chosen here:

$$x' = \frac{x - \bar{x}}{\sigma} \quad (16)$$

In mention pairs representations, the attribution feature and distance feature of the mention are added. The attribution feature is whether the two mention are in the same sentence. The features here is slightly different from the above (in span representations) in that the former is a mention-mention feature, while the latter is a single mention's own feature. So, these differences also determine that they will play different roles.

V. EXPERIMENTS

We use the Chinese coreference resolution data from the CoNLL-2012 shared task [13] in our experiments. This dataset contains 1391 training documents, 172 development documents, and 167 test documents, the domain distribution of its corpus is shown in Table 1.

Table 1: Domain distribution of the corpus

domain	Volume of corpus
News Hotline	250k
Broadcast News	250k
Broadcast Dialogue	150k
Conversational Speech	150k
Network Data	100k
English text	300

The hyperparameters of the experiments were selected as shown in Table 2. Since the hyperparameters do not change during the training of the neural network, pre-setting the appropriate hyperparameters for the model allows the model to learn effectively from the data. We conducted comparison experiments for different hyperparameters to filter out the best values. The experimental data are as follows.

Table 2: Hyperparameters settings

Name	Value	Remark
Epoch	20	Number of training iterations
BERT Learning Rate	1e-5	Learning rate of BERT parameters
Task Learning Rate	2e-4	Learning rate of other parameters
Max Segment Len	256	Maximum size of the BERT context window
Dropout Rate	0.3	Dropout scale of the model
LSTM Size	256	LSTM Hidden layer dimension
FFNN Size	1000	FFNN Hidden layer dimension

For BERT, the model achieves the best performance when the maximum size of the context window is 256, while the model performance decreases as the context window continues to increase. The analysis in this paper suggests that this is due to the fact that too long a window enhances the training difficulty and does not provide more effective contextual semantic knowledge than the original one.

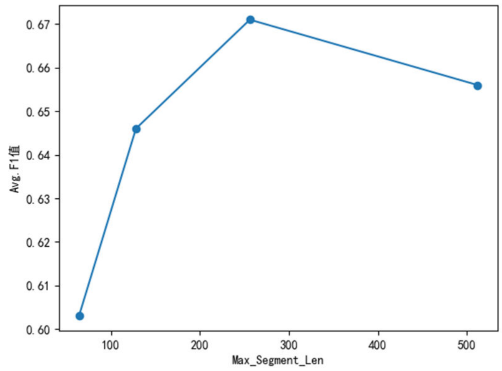


Figure 2: Max Segment Len

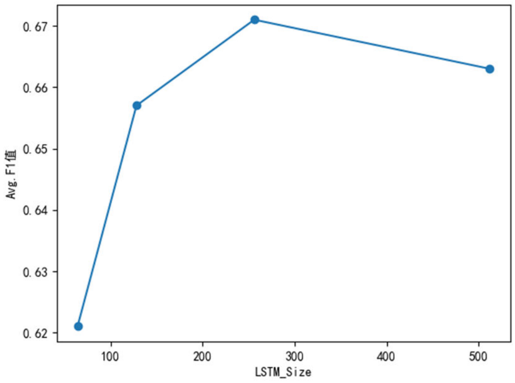


Figure 3: LSTM size

The F1 value of the model increases and then decreases as the hidden layer dimension of LSTM and FFNN increases from small to large, presumably because the model achieves the best fit when the hidden layer dimension of LSTM is 256 and the hidden layer dimension of FFNN is 1000, respectively.

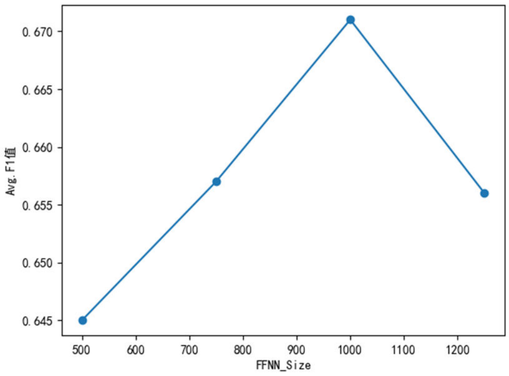


Figure 4: FFNN size

VI. Results

We report the precision, recall, and F1 for standard MUC, B-cubed, and CEAF metrics using the official CoNLL-2012 evaluation scripts. The main evaluation is the average F1 of the three metrics.

A. Baselines

Table 3 compare our model with four main baselines: (1) the original ELMo-based c2f-coref system[5], (2)its predecessor, e2e-coref[1], (3) ref-BERT[6](introduce BERT as an encoding layer to complement the contextual information of each input), and (4) c2f-BERT[14](replacing Glove and Bi-LSTM with BERT on top of c2f-coref). In particular, our model improves the state-of-the-art average F1 by 1.1. The most significant gains come from improvements in which is likely due to the fact that we have added features.

Table 3: Results on the test set on the Chinese data from the CoNLL-2012 shared task. The final column (Avg.F1) is the main evaluation metric, computed by averaging the F1 of MUV, B-cubed and CEAF.

model	MUC			B-cubed			CEAF			Avg.F1
	P	R	F1	P	R	F1	P	R	F1	
e2e-coref	68.3	62.4	65.2	58.6	50.7	54.4	52.8	50.3	51.5	57.0
c2f-coref	71.5	68.4	69.9	62.5	59.4	60.9	58.5	56.7	57.6	62.8
ref- BERT	75.6	67.9	71.5	68.6	56.3	61.8	60.4	56.2	58.2	63.9
c2f- BERT	74.9	72.4	73.6	66.4	63.3	64.8	64.1	59.8	61.9	66.8
CCRM-BERT	79.6	71.2	75.4	70.5	61.8	66.2	64.2	60.3	62.3	67.9

B. Ablations

To show the importance of each of features in our proposed model, we ablate various parts of the features and report the average F1 on the development set of the data (see Table 4).

Features play a key role in determining whether a candidate word is a mention, or whether two mentions are a mention pair. They contribute 3.5 F1 to the final result.

Table 4: Ablate various parts of the feature

	MUC			B-Cubed			CEAF			Avg.F1
	P	R	F1	P	R	F1	P	R	F1	
-all_Feature	75.4	67.9	71.5	68.1	57.5	62.5	60.8	57.4	59.1	64.3
-pOs_Feature	79.3	70.9	74.9	69.8	61.2	65.2	63.1	59.5	61.2	67.1
-dis_Feature	79.9	71.5	75.7	70.1	61.4	65.8	62.2	58.4	60.3	67.3

-wid_Feature	79.4	71.2	75.3	69.9	61.5	65.7	63.5	59.3	61.4	67.4
-Sentence_Feature	79.4	71.1	75.3	69.9	62.2	66.1	63.2	59.7	61.5	67.6
CCRM-BERT	79.6	71.2	75.4	70.5	61.8	66.2	64.2	60.3	62.3	67.9

VII. Analysis

To highlight the strengths and weaknesses of our model, we conducted two sets of experiments, and our analysis of the results of these two sets of experiments is given below.

A. Ablations

In Table 3, the c2f-coref adds higher-order prior word relations to the e2e-coref, and the F1 is improved by 5.8. The effect is significantly improved in the relatively complex context of CoNLL, which proves the usefulness of higher-order prior word relations in the model. C2f-BERT outperforms c2f-coref by 4 F1, illustrating the feasibility of BERT as an encoding layer instead of word embedding layer, and the superiority of BERT in understanding contextual information and generating the corresponding contextualized vectors. CCRM-BERT achieved better co-reference results on MUC, B-Cubed and F1, with 1.8%, 1.4% and 1.1% improvement compared to c2f-BERT, respectively, indicating that the additional contextual information of Bi-LSTM can lead to better results of the model and proving the effectiveness of CCRM-BERT model.

B. Features

In Table 4, we ablate various parts of the features. The most significant impact on the model performance is the part of speech feature, which is added to the span representation for the purpose of better determining whether the span is a mention or not. A correct mention is mostly one or more combinations of pronouns, common nouns, proper nouns, and nouns with modifiers, while conjunctions or adverbs etc. are less likely to form a correct mention, so it is necessary to insert part of speech features in span representations. The distance feature has the largest effect in MUC, which is due to the fact that MUC prefers long span and ignores individual mentions. The farther the distance between two mentions, the smaller the probability that they are a mention pair, while the disregarded distance feature can better help the model perceive farther and longer mention pairs, but the opposite is true for shorter mention pairs.

VIII. Conclusion

We present the first Chinese co-reference resolution model that incorporates lateral information into the training. Our model ensemble improves performance on the OntoNotes benchmark by 1.1 F1. We showed the importance of features to determine whether a candidate is a mention or not.

References

[1] Lee K, He L, Lewis M, et al. End-to-end neural coreference resolution[C]// Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark: EMNLP, 2017: 188-197.

[2] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]// Proceedings of Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. USA: NAACL, 2019: 4171-4186.

[3] Hobbs J R. Resolving pronoun references[J]. Lingua, 1978, 44(4): 311-338.

[4] GROSZ B, JOSHI A, WEINSTEIN S. Centering:A framework for modelling the local coherence of discourse[J]. Journal of Computational Linguistics, 1995, 21(2): 203-225.

[5] Lee K, He L, Zettlemoyer L. Higher-Order Coreference Resolution with Coarse-to-Fine Inference[C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). New Orleans, Louisiana: NAACL, 2018: 687–692.

[6] Liu F, Zettlemoyer L, Eisenstein J. The referential reader: A recurrent entity network for anaphora resolution[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Firenze, Italy: ACL, 2019: 5918-5925.

[7] Xi X F, Zhou G D. Deep Learning-based Pronoun Referential Disambiguation[J]. Acta Scientiarum Naturalium Universitatis Pekinensis,2014,50(01):100-110.

[8] Lu G Y, Wu Y J, Li R, Guan Y, Guo s r. A study of reference resolution based on the semantics of Chinese frames[J]. Computer Engineering,2020,46(10):74-80+87.

[9] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[C]// International Conference on Learning Representations Workshop Track Proceedings. Scottsdale, AZ, USA: ICLR, 2013.

[10] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Proceedings of the 31st Conference and Workshop on Neural Information Processing Systems. Long Beach, CA, USA: NIPS 2017: 5998-6008.

[11] Peters M E, Neumann M, Iyyer M, et al. Deep contextualized word representations[C]// Proceedings of Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. USA: NAACL, 2018: 2227-2237.

[12] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.

[13] Pradhan S, Moschitti A, Xue N, et al. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes[C]// Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Jeju Island, Korea: CoNLL | WS, 2012: 1-40.

[14] Joshi M, Levy O, Weld D S, et al. BERT for coreference resolution: Baselines and analysis[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing/ International Joint Conference on Natural Language Processing. Hong Kong, China: EMNLP|IJCNLP, 2019: 5802-5807.

[15] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[C]// Proceedings of the 3rd International Conference on Learning Representations. San Diego, CA, USA: 2015.

Aspect-Level Sentiment Analysis with Multi-Channel and Graph Convolutional Networks

Jiajun Wang, Xiaoge Li

Abstract**—**The purpose of the aspect-level sentiment analysis task is to identify the sentiment polarity of aspects in a sentence. Currently, most methods mainly focus on using neural networks and attention mechanisms to model the relationship between aspects and context, but they ignore the dependence of words in different ranges in the sentence, resulting in deviation when assigning relationship weight to other words other than aspect words. To solve these problems, we propose a new aspect-level sentiment analysis model that combines a multi-channel convolutional network and graph convolutional network (GCN). Firstly, the context and the degree of association between words are characterized by Long Short-Term Memory (LSTM) and self-attention mechanism. Besides, a multi-channel convolutional network is used to extract the features of words in different ranges. Finally, a convolutional graph network is used to associate the node information of the dependency tree structure. We conduct experiments on four benchmark datasets. The experimental results are compared with those of other models, which shows that our model is better and more effective.

Keywords**—**Aspect-level sentiment analysis, attention, multi-channel convolution network, graph convolution network, dependency tree

I. INTRODUCTION

THE purpose of aspect-level sentiment analysis (ASA) is to identify the sentiment of each aspect of comments in an automated way. ASA can be summarized into two tasks: (1) to extract the aspects of the target sentence; (2) to classify the sentiment of aspects. In this paper, we mainly focus on the sentiment classification task of aspects.

So far, the problem of sentiment classification has been solved by many methods. The early methods are based on dictionaries, but the construction and updating of dictionaries are very complicated and the generalization ability is poor. Support vector machine [1], decision tree [2] and Naive Bayesian [3] are also used to solve classification problems, but they are limited by the poor processing ability and generalization ability of complex texts. Due to the rise of deep learning and its powerful modeling ability, many people use neural networks to build models to establish the relationship between aspect words and context. Studies [4], [5] show that this method can strengthen the degree of association between aspect and other words. After that, more improved methods appeared on the basis of neural network, such as adding attention mechanism [6]-[9] and integrating the relative position relationship [10], the document knowledge [11] and commonsense knowledge [12] of aspects. These methods make the effect of aspects-level sentiment classification better. However, most of them ignore important features of words and word dependencies in sentences. Although the attention mechanism allows the encoder to obtain associations between aspects and other words during encoding, it is difficult to capture the dependencies between words. These problems can lead to significant deviations in giving weight to words other than aspect words in the sentence relative to aspects words.

In order to solve these problems, the dependency tree structure is introduced into the task. Dependency trees can shorten the path between aspect and evaluation words, and can capture dependencies between words. For example, "I ate the salad in this restaurant, that was too much sauce in it, so it tasted too sweet," using attention mechanism and relative position distance to associate aspect words with evaluation words, it will be difficult to connect salad with its evaluation word (sweet) for the above sentence. However, using syntactic dependency-based analysis can shorten the direct distance between salad and sweet, so as to get better results in the later modeling. Graph convolution neural network [13] can capture the local relationship of nodes in the graph by combining the characteristics of nodes. The dependency tree is used as input and encoded by bidirectional Long Short-Term Memory neural network (LSTM). The results are used as the characteristics of graph convolution neural network nodes, and then graph convolution neural network is used for aspects-level sentiment classification [14]. After that, there are more and more fusion based on dependency tree and neural network [15]-[18]. However, when encoding node features, they only focus on words and dependencies between words in the same range, lacking dependencies in different ranges. In the process of position fusion, the relative position is used instead of the word and the path between words in the dependency tree, which will give wrong weight to some irrelevant words.

In this paper, we combine multi-channel convolution and graph convolution to solve the problem of word dependence in different ranges, and also solve the optimization problem of assigning weight between words.

The main works of this paper are as follows:

We propose a new aspects-level sentiment analysis model. The multi-channel convolution neural networks and graph convolution neural networks are combined for aspects-level sentiment classification. Text convolution neural network

Jiajun Wang, NLP Laboratory, Xi'an University of Posts & Telecommunications, Xi'an, China (e-mail: Jiajunwang1999@163.com).

Xiaoge Li, NLP Laboratory, Xi'an University of Posts & Telecommunications, Xi'an, China (e-mail: xiaoge.li@gmail.com)

represents the text in the form of n-gram, so as to extract the local features. The graph convolution neural network is used to calculate the aspect words and context interactively. At the same time, the complex and irregular syntactic structure can be processed to obtain the long-distance syntactic information corresponding to the aspect words.

We replace the traditional relative position distance with the path distance of other words corresponding to aspect words in the dependency tree. This method solves the problem that the relative position distance cannot be correctly represented for complex and aspect words depend on long-distance information.

We combine multi-channel convolutional neural network and LSTM in different orders and compared them both experimentally under the same conditions. The results show that the best results are obtained by first using LSTM for context encoding and then using multi-channel convolutional neural networks to extract features in different ranges.

We compare the experimental results of different number of channels and the number of convolution kernels, and also compare the effect of different layers of GCN on the experimental results. These comparative experiments allow us to choose the optimal parameters for the experiments.

II. Related Work

Sentiment classification can be divided into text level, sentence level and aspect level. In contrast, aspect level sentiment analysis is more widely used and studied than the other two. It belongs to the fine-grained task of sentiment analysis, which aims to analyze the sentiment polarity of specific aspects in sentences. For example, “The food in this restaurant is delicious”. Through the aspect-level sentiment analysis of food, it can be concluded that its sentiment polarity is positive. For aspect-level sentiment analysis, attention mechanism and neural network are used to model in the early stage, and the following models are obtained: ATAE-LSTM [4] splice word embedding and aspect embedding, and then passes the result through attention; MEMNET [5] uses multi-layer computing layers, each layer includes an attention layer and a linear layer; IAN [8] uses the attention mechanism to model aspects and contexts; RAM [6] (Recurrent Attention Network on Memory) is used to extract sentiment information separated by long distance. After that, the basic neural network has been improved and innovated, and many new models have been obtained. PF-CNN [19] uses a new parametric convolutional neural network. MGAN [7] can learn the representation containing sentence and aspect related information, integrate it into the multi granularity sentence modeling process, and finally get a comprehensive sentence representation. TRANSCAP [20] proposes a transfer capsule network model that transforms document level knowledge into aspect-level sentiment classification. IACAPSNET [21] proposes a capsule network with Interactive Attention. Although these models have achieved good results in aspect-level sentiment analysis tasks, they do not make use of the dependency between words.

In order to model the dependency of words in aspect-level sentiment analysis, dependency tree and GCN (Graph Convolution Network) are introduced. After that, many people modeled on this basis. CDT [14] uses LSTM to represent the characteristics of a sentence, and further improves the embedded GCN to directly operate the dependency tree of the sentence. ASGCN [16] passes the features encoded by LSTM, through GCN layer, and then carries result operation with the initial coding. AEGCN [18] proposed in this paper is mainly composed of multi head attention and an improved graph convolution network based on sentence dependency tree. kumaGCN [22] integrates graph convolution neural network with gate and attention mechanism.

III. The Proposed Model

The architecture of our model (TANGCN) is shown in Fig. 1.

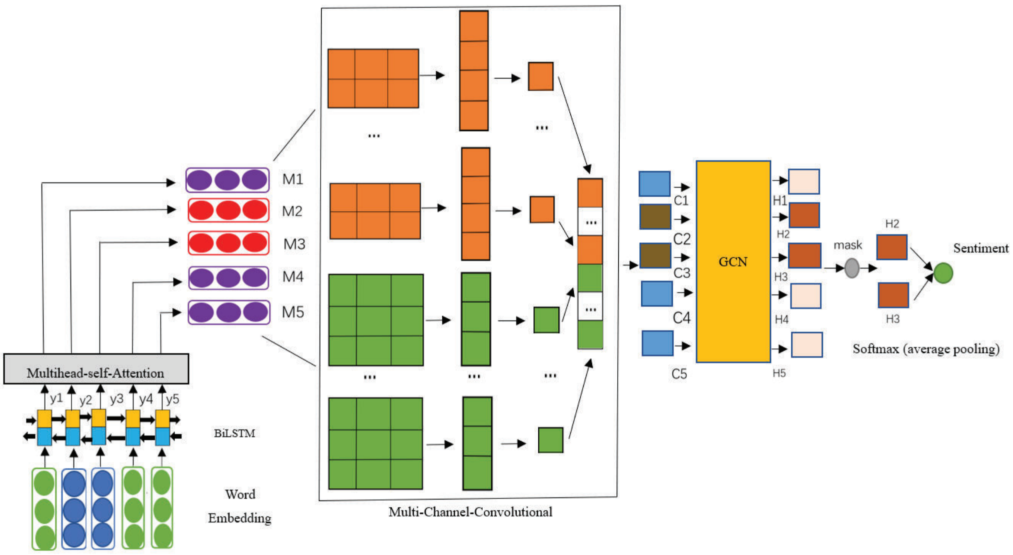


Fig. 1 Model architecture

A. Relative Graph Path of Aspects

In the past, most people used relative position to determine the positional relationship between aspects words and other words. For example, "This computer screen is very cool." The relative positional relationship between aspect (screen) and other words is shown in Fig. 2.

This computer screen looks very cool.

-2 -1 0 1 2 3 4

Fig. 2 Relative position of aspects

We propose to replace the previous relative position annotation with the path of aspects and other words in the dependency tree. The above sentence is represented in the form of a graph in the form of a dependency tree as shown in Fig. 3.

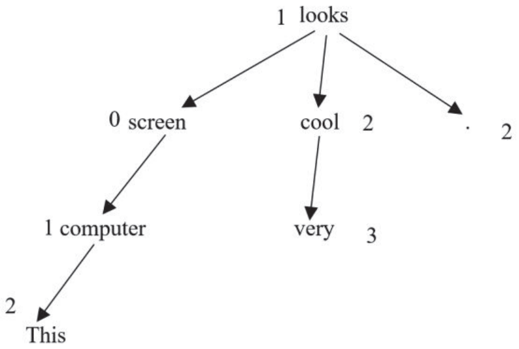


Fig. 3 Dependency tree

From the graph structure of Fig. 3, a new position code can be obtained as shown in Fig. 4. It is embedded into word vector.

This computer screen looks very cool.

2 1 0 1 3 2 2

Fig. 4 Relative graph path of aspects

B. Word Embedding

The 300-dimensional word vector of pre-trained Glove, 50-dimensional part of speech representation and 50-dimensional path representation are used to represent the input sentence. The sentence is represented by $X = \{X_1, X_2, X_3, \cdots, X_{n-1}, X_n\}$. The aspect is represented by $A = \{A_1, A_2, A_3, \cdots, A_{t-1}, A_t\}$. n is the length of the sentence and t is the length of the aspect word. The dimension of word embedding is 400.

C. Bi-LSTM Layer

LSTM can only predict the output of the next time according to the timing information of the previous time. But on most issues, we should not only pay attention to the previous information, but also pay attention to the future information, so bidirectional LSTM (BiLSTM) is introduced. The BiLSTM structure is shown in Fig. 5.

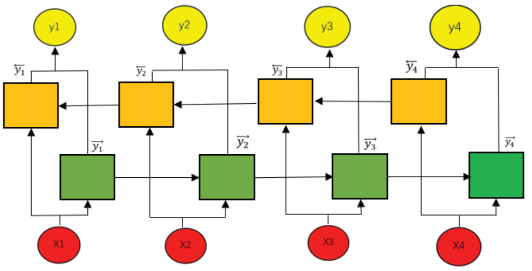


Fig. 5 BiLSTM structure

The core structure of BiLSTM can be regarded as an ordinary unidirectional LSTM, which is divided into two directions, one is forward with the input timing and the other is reverse with the input timing. As shown in Fig. 5, after the sentence $X = \{X_1, X_2, X_3, \cdots, X_{n-1}, X_n\}$ is input to BiLSTM, the forward LSTM represents the learned hidden state as $\{\overrightarrow{y_1}, \overrightarrow{y_2}, \overrightarrow{y_3}, \cdots, \overrightarrow{y_{n-1}}, \overrightarrow{y_n}\}$ and the hidden state learned by reverse LSTM is $\{\overleftarrow{y_1}, \overleftarrow{y_2}, \overleftarrow{y_3}, \cdots, \overleftarrow{y_{n-1}}, \overleftarrow{y_n}\}$. Finally, the two are spliced to get $Y = \{y_1, y_2, y_3, \cdots, y_{n-1}, y_n\}$. In this way, the context information corresponding to the aspect word is captured.

D.Multi-Head-Self-Attention Layer

Multi-head self attention is an attention mechanism that can operate in parallel in space. In this paper, $Y = \{y_1, y_2, y_3, \cdots, y_{n-1}, y_n\}$ is regarded as an input feature. After passing through the attention layer, a new feature $M = \{M_1, M_2, M_3, \cdots, M_{n-1}, M_n\}$ containing the degree of association between words is obtained. Firstly, the input features Y are multiplied by three matrices (B1, B2 and B2) to obtain three matrices (Q, K and V).

$$Q = YB1 \tag{1}$$

$$K = YB2 \tag{2}$$

$$V = YB3 \tag{3}$$

Secondly, the definition of attention is given.

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

The above formula is actually the weighted calculation of V. The weight is softmax (*) before V. In softmax, the similarity score is obtained by the dot product operation of Q and K, and then use $\sqrt{d_k}$ to adjust size. The multi-head-self attention mechanism is obtained by linearly changing the initial Q, K and V matrix to obtain the following matrix.

$$Q = \{Q1, Q2, Q3, \cdots, Q_n\} \tag{5}$$

$$K = \{K1, K2, K3, \cdots, K_n\} \tag{6}$$

$$V = \{V1, V2, V3, \cdots, V_n\} \tag{7}$$

The h is the number of heads. Each head does not share a parameter matrix. Make an attention to each head, repeat h times, and then splice the results.

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \quad (8)$$

$$MultiHead(Q, K, V) = Concat(head_1, \cdots, head_h) \quad (9)$$

E. Multi-Channel-Convolutional Layer

We use two types of convolution kernels in the convolutional layer to extract the associations between words in different scopes. Their width is equal to the dimension of the input word vector. Their heights are 2 and 3 respectively. The number of convolution kernels is 150. In this paper, $M = \{M_1, M_2, M_3, \cdots, M_{n-1}, M_n\}$ is regarded as an input feature. After passing through this layer, a new feature $C = \{C_1, C_2, C_3, \cdots, C_{n-1}, C_n\}$ is obtained. Multi-Channel-Convolutional structure is shown in Fig. 6.

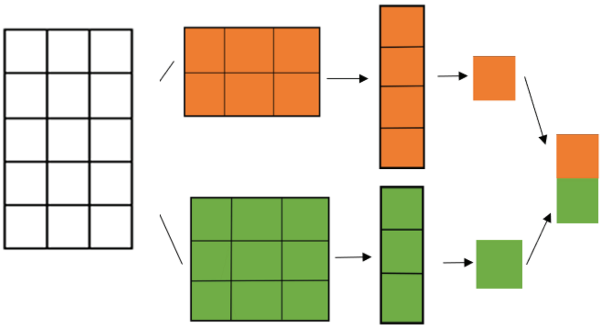


Fig. 6 Multi-Channels-Convolutional structure

Set the number of input features be n. The input matrix is $M_{[n\times word\ embedding_{dim}]} \cdot A_{[i:j]}$ stands for lines i to j. The convolution operation can be expressed by the following formula; w is convolution kernel; h is the height of the convolution kernel.

$$o_i = w \cdot A_{[i:i+h-1]}, i = 1,2, \ldots, n-h+1 \quad (10)$$

The obtained result is added with bias b, and then activated using the activation function to obtain the desired feature. Finally, max pooling is used for pooling.

$$c_i = f_{(o_i+b)} \quad (11)$$

F. Graph Convolution Network Layer

Graph Convolution Network is suitable for processing graph structure data with rich correlation information. According to the feature representation $C = \{C_1, C_2, C_3, \cdots, C_{n-1}, C_n\}$ of the previously generated words and the dependency tree generated according to the sentences, the information required by the graph convolution neural network can be obtained. The dependency tree can be seen as a graph $G = (\mathrm{V}, \mathrm{E})$. Nodes V represent words, and edges E represent the relationship between words. $C = \{C_1, C_2, C_3, \cdots, C_{n-1}, C_n\}$ represents the characteristic representation of each word. By constructing $|V| \times |V|$ the adjacency matrix A can be obtained. $A_{\mathrm{ij}} = 1$ if node i is connected to node j, otherwise $A_{\mathrm{ij}} = 0$. In order to make the GCN model and embed nodes more effectively, each node in the graph is allowed to have self-loops.

$$\tilde{A} = A + I \quad (12)$$

Where $\tilde{A}$ is the adjacency matrix plus the identity matrix I of the graph. The graph convolution of a node can be described as:

$$\alpha^i = (\textstyle\sum_{j=1}^{n} \tilde{A}_{ij})^{-1} \quad (13)$$

$$h_i^{(k+1)} = \sigma\left(\textstyle\sum_{j=1}^{n} \alpha^i \tilde{A}_{ij}(W^k h_j^{(k)} + b^{(k)})\right) \quad (14)$$

Where W^k is the weight matrix, $b^{(k)}$ is the offset vector, σ is a nonlinear function. $h_j^{(k)}$is the hidden state of node j after passing through the k-1-layer GCN. α^i is the reciprocal of the degree of node i in the graph. After passing through the k-layer GCN, we get the final output $H = \{h_1, h_2, h_3, \cdots, h_{n-1}, h_n\}$ of the k-layer. Then I mask it to get the feature representation $A = \{a_1, a_2, a_3, \cdots, a_{t-1}, a_t\}$ of the aspect word we need. Where t is the length of aspect word. Mask function is an operation function that multiplies input and mask matrix.

$$\{a_1, a_2, \cdots, a_{t-1}, a_t\} = MASK(\{h_1, h_2, \cdots, h_{n-1}, h_n\}) \quad (15)$$

This feature A represents the fusion of context related information extracted by LSTM, the degree of association between words obtained by attention mechanism, the feature relationship between words in different ranges extracted by multi-channel convolution and the information aggregated by GCN. We use average pooling to average the information in the aspect word vector to obtain the final feature representation.

$$L = Average\ pooling(\{a_1, a_2, a_3, \cdots, a_{t-1}, a_t\}) \quad (16)$$

Finally, the feature representation is input to the softmax layer for sentiment distribution probability calculation. Where r is the category of classification. W_p and b_p are the learned weight matrix and bias, respectively.

$$S = Softmax\left(W_p L + b_p\right) \quad (17)$$

IV. Model Training

This model is trained by the standard gradient descent algorithm with the cross-entropy loss.

$$Loss = -\textstyle\sum_i \sum_{j\in J} y_i^j log\ \hat{y}_i^j \quad (18)$$

Where i represents the subscript of the i-th aspect sentence and j represents the sentiment category of the j-th sample sentence. y represents the true sentiment distribution of the sentence. $\hat{y}$ represents the predicted sentiment distribution of the sentence.

V. EXPERIMENTS

A. Datasets

In order to prove the superiority and effectiveness of TANGCN, we evaluate the performance of this model on restaurant reviews (Rest14; Rest16[23]), laptop reviews (Laptop14) [24] and Twitter reviews [25]. The details of the experimental data are shown in the Table I.

TABLE I
DISTRIBUTION OF SAMPLES BY CLASS LABELS ON BENCHMARK DATASETS

Polarity	Dataset	Rest14	Laptop14	Rest16	Twitter
Positive	Train	2164	976	1657	1507
	Test	727	337	611	172
Neutral	Train	637	455	101	3016
	Test	196	167	44	336
Negative	Train	807	851	748	1528
	Test	196	128	204	169

B. Parameters

In this experiment, we use a 96-dimensional LSTM embedding for each word. The dependency tree structure of sentences is developed by Stanford parser. The batch size is 32. The number of self-multi-head attention heads is 8. Multi-channel convolution uses two convolution kernels, all of which have widths equal to the dimension of the word vector, and heights of 2 and 3 respectively. The number of convolution kernels is 150.The number of GCN layers is 2.

C. Effect of CNN and LSTM order

Multi-channel-Convolutional-LSTM: Firstly, the local features of the text are extracted by Multi-Channel-Convolutional, and then the long-distance features of these local features are extracted by LSTM.

LSTM-Multi-Channel-Convolutional: Firstly, the long-distance features of the text are extracted by LSTM to obtain the new text fused with the context, and then the local features of the new text are extracted by Multi-Channel-Convolutional. The results on rest16 and twitter are shown in Fig. 7 and Fig. 8

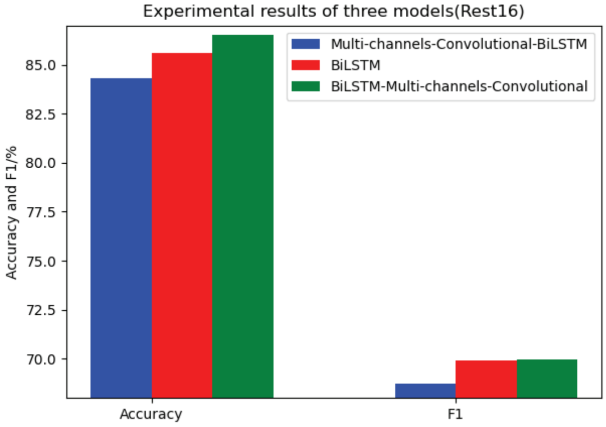


Fig. 7 Model effect comparison (Rest16)

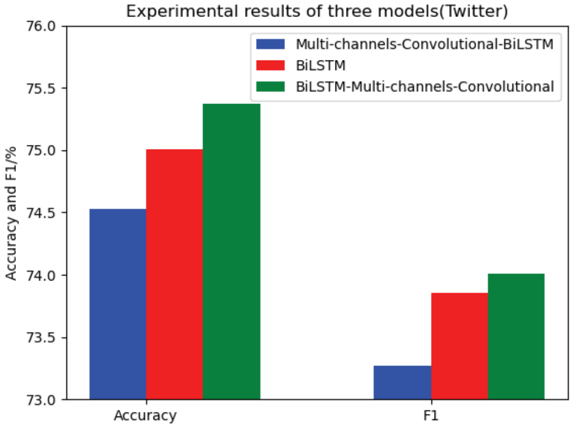


Fig. 8 Model effect comparison (Twitter)

From the experimental data, it can be concluded that the effect of LSTM-Multi-Channel-Convolutional is the best and Multi-Channel-Convolutional-LSTM is the worst.

D. Effect of Different Convolution Kernel Size and Number

In the process of text convolution using multi-channel convolutional networks, the parameter that has a great impact on the size and number of convolution kernels. In this paper, the experimental results of multi-channel convolution kernel and single channel convolution are compared. We use the width of the convolution kernel equal to the dimension of the word vector. We use convolution kernels with heights of 2, 3 and 4 to combine to obtain two multi-channel convolutional layers. One combination is convolution kernel heights 2 and 3 ([2,3]), and the other is convolution kernel heights 2, 3 and 4 ([2,3,4]). For single-channel convolutional layers, we use a convolutional kernel height of 2 ([2]). The experimental data are Twitter and Rest14. In order to compare the effect of different numbers of convolution kernels on our experimental results, in the case of a combination of kernel heights of 2 and 3 ([2,3]), we conduct experimental comparisons with the number of convolution kernels of 50, 100, 150 and 200 respectively. The experimental results are shown in Fig. 9 and Fig. 10.

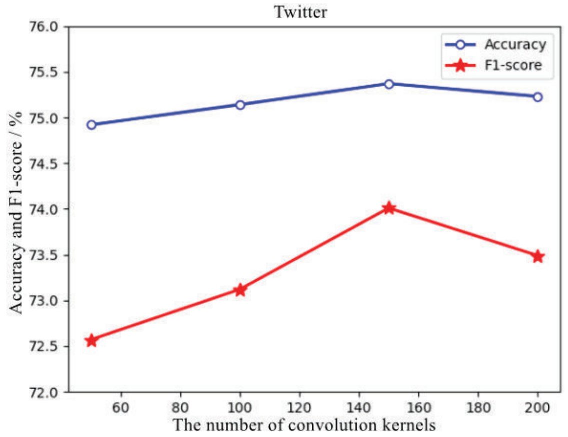


Fig. 9 Effect of different numbers of convolution kernels (Twitter)

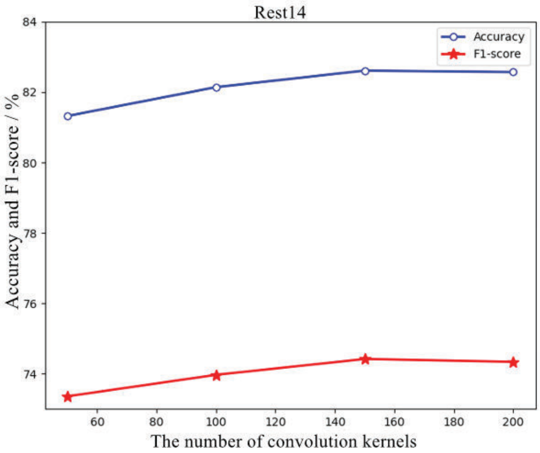


Fig. 10 Effect of different numbers of convolution kernels (Rest14)

It can be seen from the Fig. 9 and Fig. 10 that under the same number of channels, the best result is to choose 150 convolution kernels for the four types. In order to compare the effect of different size of convolution kernels on our experimental results, in the case of using 150 convolution kernels, we conduct experimental comparisons with the size of convolution kernels of [2], [2,3] and [2,3,4] respectively. The experimental results are shown in Fig. 11 and Fig. 12.

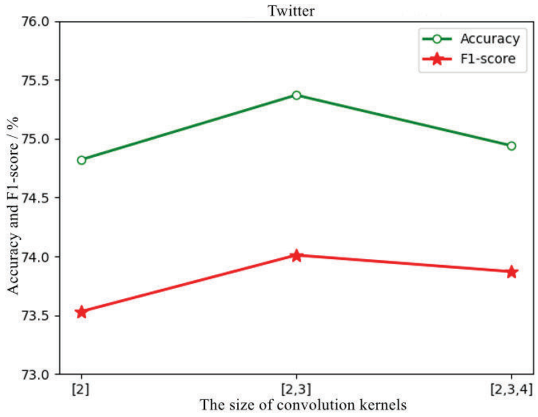


Fig. 11 Effect of different convolution kernel sizes (Twitter)

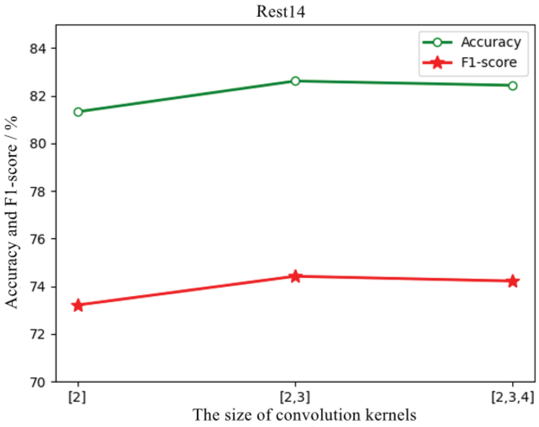


Fig. 12 Effect of different convolution kernel sizes (Rest14)

It can be seen from the Fig .11 and Fig .12 that [2,3] is the best for the three channel numbers under the same number of convolution kernels

E. Effect of GCN Layers

The number of layers in graph convolution neural network is a parameter that has a great influence on the results. Based on the original model, the number of layers of GCN is changed, and the experimental results of Laptop14 are statistically analyzed. The comparison results are shown in the Fig. 13.

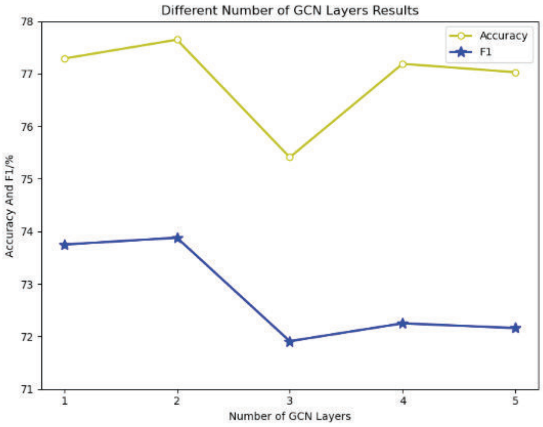


Fig. 13 Effects of the number of GCN layers

It can be seen from the data statistical chart that with the increase of GCN layers, the accuracy of experimental results and F1-Score are changing. Firstly, after increasing from 1 layer to 2 layers, the accuracy and F1-Score increase. Then from 2 layers, the accuracy and F1-Score show a downward trend with the increase of the number of layers. Therefore, the GCN in this model adopts a 2-layer structure.

F. Experimental Results

TableII shows the experimental results of each model on four different data sets. Compared with the model after integrating the dependency tree structure and considering the dependency relationship between words, we can find that the experimental results obtained by using only attention mechanism and neural network for aspect words and context modeling in the early stage are lower, Thus, it is proved that the dependency tree can improve the aspect-level sentiment analysis, and the graph convolution neural network is effective for this task. Compared with the models integrated with the dependency tree, it can be seen that the experimental results of CDT, ASGCN, AEGCN and kumaGCN are basically lower than the TANGCN model in this paper in accuracy and F1-Score, and only the accuracy of ASGCN on rest16 dataset is higher than the model in this paper. Therefore, it is proved that using the path distance of graph as the coding fusion of words and using multi-channel text convolution to extract the dependencies of words in different ranges can make the classification task more accurate.

TABLE II
EXPERIMENTAL RESULTS

Models	Rest14		Rest16		Twitter		Lap14	
	ACC	F1	ACC	F1	ACC	F1	ACC	F1
ATAE-LSTM	77.20	-	-	-	-	-	68.70	-
MEMNET	80.95	-	-	-	-	-	72.21	-
IAN	78.60	-	78.60	-	-	-	72.10	-
RAM	80.23	70.80	-	-	69.36	67.30	74.49	71.35
PF-CNN	79.20	-	-	-	-	-	70.06	-
MGAN	81.25	71.94	-	-	72.54	70.81	75.39	72.47
HSCN	77.80	70.20	-	-	69.60	66.10	76.10	72.50
TRANSCAP	79.55	71.41	-	-	-	-	73.87	70.10
IACAPSNET	81.79	73.40	-	-	-	-	76.80	73.29
ANTM	82.49	72.10	-	-	72.35	69.45	75.84	72.49
CDT	82.30	74.02	85.58	69.93	74.66	73.66	77.19	72.99
ASGCN	81.22	72.94	**88.99**	67.48	72.69	70.59	75.55	71.05
AEGCN	81.04	71.32	87.39	68.22	73.16	71.82	75.91	71.63
kumaGCN	81.43	73.64	-	-	72.45	70.77	76.12	72.42
TANGCN	**82.61**	**74.42**	86.51	**69.97**	**75.37**	**74.01**	**77.65**	**73.88**

VI. CONCLUSION

In this paper, we use dependency trees in aspect-level sentiment analysis to enforce semantic dependencies between words and address long-distance dependencies. We use a multi-channel convolutional neural network to solve the problem of extracting dependencies between words in different ranges. GCN is used to aggregate the information around each word and finally get the sentiment polarity of the aspect word. Compared with other models, our model is more accurate and effective on multiple datasets. In future work, we will pay more attention to the research and improvement of graph structure construction to solve the noise problem of dependency trees.

REFERENCES

[1] E. Burnaey, D. Smolyakoy, “One-Class SVM with Privileged Information and its Application to Malware Detection,” *IEEE* (2016).

[2] L. H. Zhang, et al. “Factors affecting decision tree classification method over TM image.” in *Forest Research, Beijing*, pp. 1-5, 2014.

[3] S. Mishra, M. Panda, “Histogram of oriented gradients-based digit classification using naive Bayesian classifier.” in *Progress in Computing, Analytics and Networking*, Springer, Singapore, pp. 285-294.

[4] Y. Wang, M. Huang, X. Zhu, and L. Zhao, “Attention-based LSTM for aspect-level sentiment classification,” in *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 606-615, Nov 2016.

[5] D. Tang, B. Qin, X. Feng, and T. Liu, “Effective LSTMs for Target-Dependent Sentiment Classification.” in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 3298–3307, 2016.

[6] Sun, L. Bing, and W. Yang, “Recurrent attention network on memory for aspect sentiment analysis.” in *Proceedings of the 2017 conference on empirical methods in natural language processing*, pp. 452-461, Sep 2017.

[7] F. Fan, Y. Feng, and D. Zhao, “Multi-grained attention network for aspect-level sentiment classification.” in *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 3433-3442, 2018.

[8] D. Ma, S. Li, X. Zhang, and H. Wang, “Interactive attention networks for aspect-level sentiment classification.” in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pp. 4068–4074, 2017.

[9] R. He, W. S. Lee, H. T. Ng, and D. Dahlmeier, “Effective attention modeling for aspect-level sentiment classification.” in *Proceedings of the 27th international conference on computational linguistics*, pp. 1121-1131, Aug 2018.

[10] S. Gu, Zhang, L. Zhang, Y. Hou, and Y. Song, “A position-aware bidirectional attention network for aspect-level sentiment analysis.” in *Proceedings of the 27th international conference on computational linguistics*, pp. 774-784, Aug 2018.

[11] R. He, W. S. Lee, H. T. Ng, and D. Dahlmeier, “Exploiting document knowledge for aspect-level sentiment classification.” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2018.

[12] Y. Ma, H. Peng, and E. Cambria, "Targeted Aspect-Based Sentiment Analysis via Embedding Commonsense Knowledge into an Attentive LSTM.” *AAAI* 2018.

[13] T. N. Kip F, M. Welling, “Semi-supervised classification with graph convolutional networks.” 2016.

[14] K. Sun, Zhang, R. Zhang, S. Mensah, Y. Mao, and X. Liu, “Aspect-level sentiment analysis via convolution over dependency tree.” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5679-5688, Nov 2019.

[15] B. Huang, K. M. Carley, “Syntax-aware aspect level sentiment classification with graph attention networks.” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5472–5480 2019.

[16] C. Zhang, Q. Li, D. Song, “Aspect-based Sentiment Classification with Aspect-specific Graph Convolutional Networks.” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4560-4570, 2019.

[17] K. Wang, W. Shen, Y. Yang, X. Quan, and R. Wang, “Relational Graph Attention Network for Aspect-based Sentiment Analysis.” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3229–3238, Online 2020.

[18] L. Xiao, X. Hu, Y. Chen, Y. Xue, and T. Zhang, “Targeted sentiment classification based on attentional encoding and graph convolutional networks.” Applied Sciences, 10(3), 957, 2020.

[19] B. Huang, K.M. Carley, “Parameterized convolutional neural networks for aspect level sentiment classification” 2019.

[20] Z. Chen, T. Qian, “Transfer capsule network for aspect level sentiment classification.” in *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 547-556, July 2019.

[21] C. Du, H. Sun, J. Wang, Q. Qi, and M. Liu, “Capsule Network with Interactive Attention for Aspect-Level Sentiment Classification.” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

[22] C. Chen, Z. Teng, and Y. Zhang, “Inducing target-specific latent structures for aspect sentiment classification.” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online. Association for Computational Linguistics (2020)*, pp. 5596–5607, 2020.

[23] M. Pontiki, et al., “SemEval-2016 Task 5: Aspect Based Sentiment Analysis.” in *International Workshop on Semantic Evaluation 2018*.

[24] M. Pontiki, et al., “SemEval-2014 Task 4: Aspect Based Sentiment Analysis. in *Proceedings of the 8th International Workshop on Semantic Evaluation(SemEval-2014)*.

[25] L. Dong, F. Wei, C. Tan, D. Tang, M. Zhou, K. Xu, “Adaptive recursive neural network for target-dependent twitter sentiment classification.” In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 49–54. Baltimore, MD, USA, 23–25 Jun 2014.

Accurate Positioning Method of Indoor Plastering Robotbased on Line Laser

Guanqiao Wang, Hongyang Yu

Abstract—There is a lot of repetitive work in the traditional construction industry. These repetitive tasks can significantly improve production efficiency by replacing manual tasks with robots. Therefore, robots appear more and more frequently in the construction industry. Navigation and positioning is a very important task for construction robots, and the requirements for accuracy of positioning are very high. Traditional indoor robots mainly use radio frequency or vision methods for positioning. Compared with ordinary robots, the indoor plastering robot needs to be positioned closer to the wall for wall plastering, so the requirements for construction positioning accuracy are higher, and the traditional navigation positioning method has a large error, which will cause the robot to move. Without the exact position, the wall cannot be plastered or the error of plastering the wall is large. A new positioning method is proposed, which is assisted by line lasers and uses image processing-based positioning to perform more accurate positioning on the traditional positioning work. In actual work, filter, edge detection, Hough transform and other operations are performed on the images captured by the camera. Each time the position of the laser line is found, it is compared with the standard value, and the position of the robot is moved or rotated to complete the positioning work. The experimental results show that the actual positioning error is reduced to less than 0.5 mm by this accurate positioning method.

Keywords—Indoor plastering robot, Navigation, Precise positioning, Line Laser, Image processing

I. Introduction

As the population growth rate continues to decrease and the labor force decreases, all walks of life seek to improve productivity through the widespread use of robots, but due to the uncertainty of the construction industry, the complex construction environment, and the degree of danger is second only to the mining industry, the current construction industry is still dominated by traditional worker decoration, which makes it difficult to improve production efficiency.

The interior decoration work is also gradually replaced by robots. The main tasks that can be done by robots in interior decoration are as follows: Article [1] introduces the paste of tiles, and article [2] introduces the use of robots to complete the painting of walls, article [3, 4] introduced the plastering work of the wall. The plastering work of the wall means that for the newly built house, it is often necessary to manually plaster the wall with cement mortar, so that the wall of the whole room becomes flat, which is more conducive to the subsequent fine decoration work.

For indoor plastering, robot positioning and navigation technology is very important. In recent years, many indoor positioning technologies have emerged, such as base station positioning, wi-fi positioning, radio frequency tag positioning, visual positioning and other technologies [5–7]. At present, the commonly used robot positioning and navigation technologies are mostly simultaneous localization and mapping (SLAM), generally divided into lidar-based SLAM technology or computer vision technology-based SLAM technology [8], these two technologies are mostly used in sweeping robots, shopping guide robots in shopping malls and other fields.

In the wall plastering work, the robot needs to achieve autonomous navigation and positioning to the specific position of the wall to be plastered. Traditional positioning methods are generally not used in the field of construction, and SLAM methods based on vision or lidar cannot meet the requirements of building construction scenarios in terms of time and accuracy. The basic technical method of SLAM is: first, the robot scans indoors, and then according to the collected images, the feature points are matched to estimate the motion trajectory, and then the key frames are selected to finally realize the map generation. That is to say, the strategy of mapping first and then positioning is adopted. For the traditional SLAM method, the method of ”mapping first, then positioning” is adopted. Most of the time will be spent on the generation of the map, and the accuracy will also be lost on each key frame selection step.

For fine wall plastering work, the precision of the robot is higher. In the actual construction project, the flatness error of the plastered wall surface is required to be within 0.5 mm. Therefore, it is necessary to propose a method according to the actual engineering needs. New low-cost, high-accuracy, and low-time-consuming methods to implement indoor localization methods.

In order to solve the above problems, this paper proposes an accurate positioning method for indoor plastering robots based on line laser assistance. This method makes more accurate positioning and navigation of the robot by visually recognizing the laser line and adjusting it dynamically until the error is less than the threshold. Calibration ensures that the maximum accuracy of wall plastering can be achieved, so that the final plastered wall is flat.

II. Plastering Robot Structure

The structure of the entire indoor robot accurate positioning system can be simplified as shown in Fig. 1, including a line laser transmitter and the robot itself. In indoor architectural scenes, there are mainly wired laser transmitters and the robot itself. The line laser transmitter emits line lasers parallel to the wall. The robot includes a fuselage, a mechanical pole, a plastering head, a mobile car, two RGB high-definition cameras and the light receiving plate, the body is placed on the mobile trolley, the plastering head is connected to the body through a mechanical rod, the light receiving plate is set at the bottom of the back of the fuselage, and is used to capture the image on the light receiving plate. Fig. 2 shows the light

Wang Guanqiao is with the University of Electronic Science and Technology of China, China (e-mail: 812844840@qq.com).

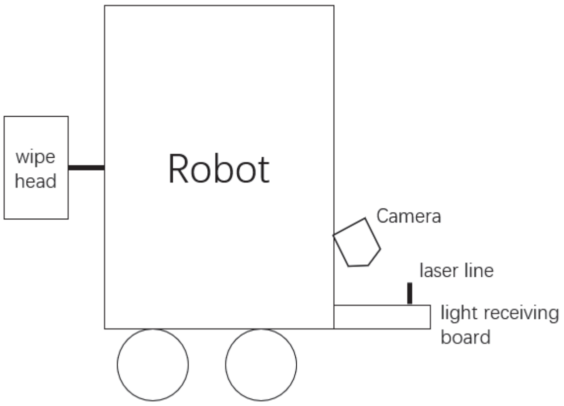


Fig. 1. Plastering Robot Structure

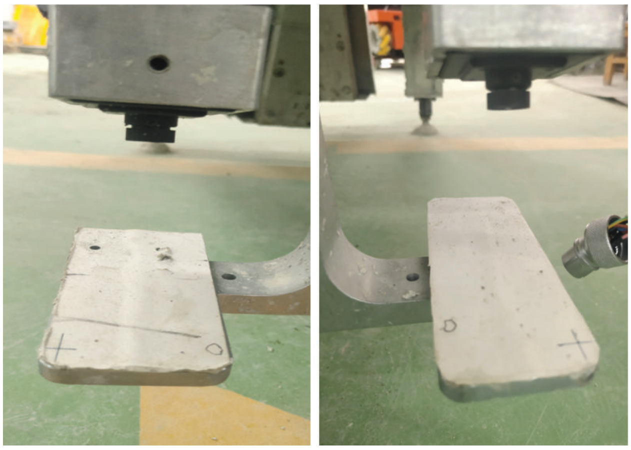


Fig. 2. The relationship between the camera and the light receiving board

receiving plate and the camera in the the actual position in the robot.

Fig. 3 shows the relationship between the robot and the laser. A line of laser line parallel to the wiper head is drawn, and the position of the line laser in the field of view of the left and right cameras is recorded to obtain the calibration values Y1 and Y2.

III. Accurate Positioning System Process

The accurate positioning of the robot can be achieved by the method based on image processing, which can make the positioning and navigation of the robot more accurate. The overall accurate positioning process can be divided into three parts: the first part is the calibration module, which is used to determine the calibration line; the second part is the image processing module, which is used to identify the laser line in real time during the accurate positioning process; and the third part is the movement parameter calculation module, which is used to control the angle and distance of the robot's movement each time. The three parts coordinate and cooperate to form the accurate positioning system of the entire robot. The entire process is shown in Fig. 4 below.

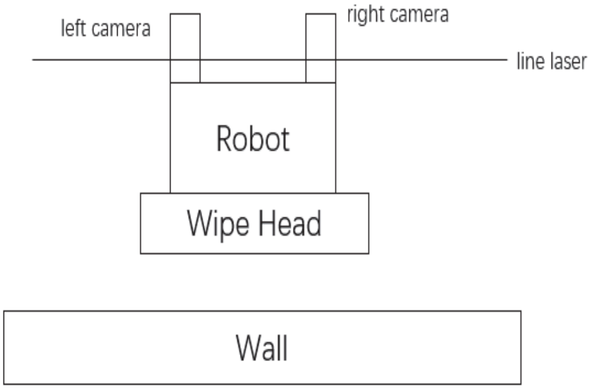


Fig. 3. The relationship between the robot and the lase

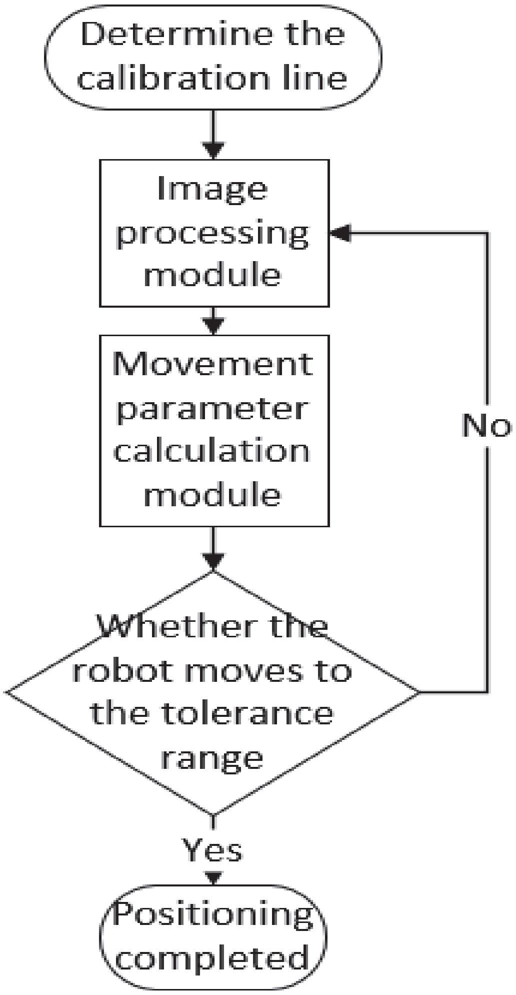

Determine the calibration line
Image processing module
Movement parameter calculation module
Whether the robot moves to the tolerance range
No
Yes
Positioning completed

Fig. 4. Accurate positioning system process

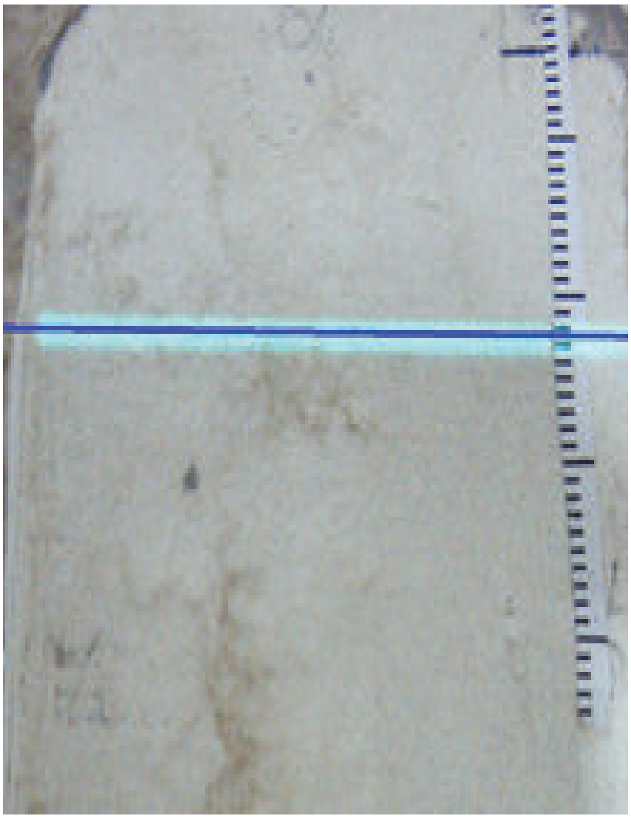


Fig. 5. Convert the laser line to a straight line

IV. Image Processing Module of Robot

In the third chapter, it is mentioned that the camera can identify the laser image to obtain specific calibration values. Taking a high-definition 1080P non-distortion camera as an example, the pixels of the captured picture are 1920*1080. After ordinary positioning and navigation, the robot can be moved to the light receiving board. To obtain the position of the line laser, for the laser line appearing on the light receiving plate, the monocular high-definition camera scans to obtain the image, as shown in Fig. 5, it is hoped that the center of the laser can be extracted as a straight line through image processing.

In order to process the laser line, it is necessary to identify the significant part of the laser, and then convert it into a straight line in the middle band, and operate through a series of image processing methods. The main process is shown in Fig. 6.

A. Histogram Equilibrium

In the actual construction environment, due to the different lighting conditions of the construction room, in the image content to be extracted, the inconsistent brightness of the image will lead to inconsistent brightness of the laser. In order to eliminate this effect, the image is converted to grayscale at the beginning. Then perform histogram equalization, which can reduce the interference caused by too high or too dark brightness, and achieve the effect of image enhancement.

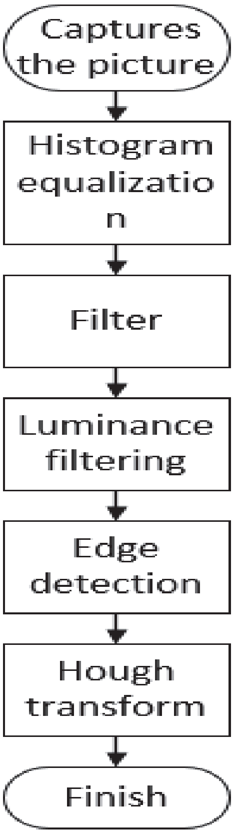


Fig. 6. Image processing flow

B. Filter

The clearer the image, the more noise it contains. The picture becomes clearer after histogram equalization. Therefore, Gaussian filtering is used to filter out the high-frequency noise in the image, and because the laser line to be identified in the image is in the horizontal direction. According to the filtering method proposed in article [9], a smaller Gaussian kernel can be used in the horizontal direction, and a larger Gaussian kernel can be used in the vertical direction. The size of the Gaussian kernel used here is (2k+1, 4k+1), after filtering, the value F_{ij} of the Gaussian filter at the original image (i, j) can be obtained by the formula (1), and then the median filter is used to eliminate the salt and pepper noise.

$$F_{ij} = \frac{1}{2\pi\sigma^2} exp(-\frac{(i-k)^2 + (j-2k)^2}{2\sigma^2}) \tag{1}$$

C. Brightness filtering

Since the scene is a construction industry scene, there will inevitably be a lot of dust particles in the image. In this case, in order to filter it, a filtering method based on the brightness feature of the laser line is proposed, which converts the image from RGB to HSV means, and then filter the HSV channel of the pixel to the V channel (brightness) to determine the brightness value of each pixel. If the brightness value exceeds the threshold, it is considered to be a brighter laser spot, otherwise it is a non-laser spot, that is, formula (2), where α, β, γ are the HSV channel thresholds set in advance through measurement and judgment. For pixels with insufficient brightness, they are directly set to full black, otherwise the pixels are still the original value.

$$P_{hsv(x,y)} \leq \alpha, \beta, \gamma \ , \ P_{hsv(x,y)} = 0 \tag{2}$$

D. Edge detection

For the filtered pixel, it is still a blurred laser strip. In order to extract the laser into a straight line, the edge of the laser needs to be extracted. The Sobel operator in the Y direction [9] is used to detect the edge of the image. Then two obvious horizontal straight lines can be obtained, which are the upper and lower edge lines of the laser.

E. Hough transform for straight line fitting

After edge detection, many straight line segments are detected on the edge, but the lengths and slopes of these straight line segments are different, and effective line segments need to be extracted. Here, the Hough transform method with transformable parameters is used, and the threshold value is selected by the voting mechanism. For many short straight line segments with many points, this line segment is represented by the starting point and the end point, and a feedback mechanism is added at the same time. If the number of endpoints is too small at this time, the parameter threshold of the Hough transform should be relaxed, otherwise, the adjustment threshold can be strengthened. After obtaining many straight line segment endpoints, first use the RANSAC algorithm [10] to filter out the outer points, and then use the least squares method to fit valid points. A large number of points on the upper and lower edges of the laser are fitted as a straight line in the middle of the laser. This method solves the problem of precision loss caused by too thick laser lines.

V. Movement Parameter Calculation Module of Robot

In the second chapter, a straight line that can represent the laser line is obtained through image processing. Since the distance between the plastering robot and the wall must be a fixed value, the position where the laser line should be can be determined by pre-calibration. This straight line is called the calibration straight line. By comparing the straight line obtained by fitting each time with the calibration straight line, the distance and angle that the robot should move can be calculated.

A straight line reflecting the actual robot position can be obtained by image processing. The intersection of the straight line and the left and right field of view of the camera is represented by Y1 and Y2. We use Y1 and Y2 to represent the actual calibration position. The resolution of the camera used is 1920*1080. Therefore, the left and right endpoints of the calibration line and the fitted line are (0, Y1), (0, Y2), (1080, $Y1\prime$), (1080, $Y2\prime$) . To enable the robot to move to the calibrated position, the translation is to move the two lines to coincide, as shown in Fig. 7.

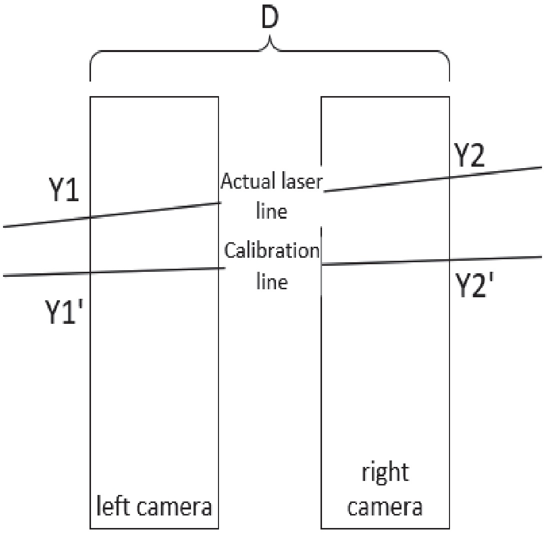


Fig. 7. Calibration line and actual straight line

TABLE I
Adjustment value each time during a accurate positioning process

Number	Distance(mm)	Angle(°)
1	8.5	0.30
2	1.3	0.06
3	0.4	0.00
4	0.1	0.00
5	0.0	0.00

In order to overlap the two straight lines, it is only necessary to calculate the distance and angle that the straight line needs to move. Through the calculation, the distance that the plastering robot needs to move in the Y direction is Y_d. The calculation formula is as formula (3), where alpha is The amount of pixels represented by one millimeter in practice, the rotation angle that the robot needs to move is theta, and the calculation formula is as formula (4), where D is the distance between the edges of the two cameras, that is, the distance marked in Fig. 7.

$$Y_d = \frac{(Y1 - Y1' + Y2 - Y2')}{2\alpha} \tag{3}$$

$$\theta = arctan\frac{(Y1 - Y1') - (Y2 - Y2')}{2\alpha D} \tag{4}$$

VI. Experiment

Based on this method, the robot is tested for accurate positioning. Table I represents the accurate positioning results in one experiment. Through five continuous accurate positioning, the distance and angle adjusted each time, it can be seen that with the increase of the number of times, each time Both the distance and the angle of movement are gradually reduced until a set threshold is reached.

Based on this method, within a distance of 2 mm, through ten times of precise positioning, the actual distance measured

TABLE II
Ten precise positioning of the actual distance value

Locating Number	precision(mm)
1	2.1
2	1.5
3	2.3
4	1.8
5	2.4
6	1.6
7	1.8
8	2.1
9	2.6

TABLE III
Indoor positioning technology comparison

Techology	precision
Wifi-based fingerprint method	1–5m
Radio Frequency Identification	0.05–5m
Ultra Wideband Technology	6–10cm
Infrared technology	5–10m
Ultrasonic technology	1–10cm
Visual positioning	0.01–1m
Inertial navigation	2–10m
Lidar SLAM	5–10mm
Accurate positioning based on line laser	0.1–0.5mm

each time is shown in Table II. It can be seen that through this method, the error of the moving distance is basically controlled within 0.5 mm.

Through this method, the accuracy of indoor positioning is improved to within the range of 0.5 mm. Compared with the traditional indoor positioning technology, the accuracy is shown in Table III. It can be seen that the accuracy of this method is higher than that of the traditional method.

VII. Conclusion

In this paper, a accurate positioning method based on line laser assistance is proposed, which helps the construction plastering robot to perform accurate positioning work by identifying the laser line through image processing, and improves the positioning accuracy to within 0.5 mm, which greatly improves the positioning accuracy of industrial robots, also makes the wall surface smoother. Has huge application value in industry. In the follow-up experiments, it is planned to use a better calculation method to abandon the disadvantage of determining the benchmark in advance, and to perform the positioning work in real time during the work process.

References

[1] King N, Bechthold M, Kane A, et al. Robotic Tile Placement: Tools, Techniques and Feasibility. Automation in Construction **39**(01), 161–166(2014)

[2] Elashry K, Glynn R. An Approach to Automated Construction Using Adaptive Programing. Robotic Fabrication in Architecture, Art and Design 2014. AnnArbor, Michigan : Springer : 51–66(2014)

[3] Asadi E, Li B, Chen I-M. Pictobot. IEEE Robotic and Automation Magazine, **25**(2), 82–94(2018)

[4] Yan R-J, Kayacan E, Chen I-M, et al. QuicaBot: Quality Inspection and Assessment Robot. IEEE Transactions on Automation Science and Engineering, **01**(99) : 1–12(2018)

[5] Dammann A, Sand S, Raulefs R. On the benefit of observing signals of opportunity in mobile radio positioning. SCC 2013; 9th International ITG Conference on Systems, Communication and Coding. VDE, 2013: 1-6(2013)

[6] Bhatt D, Babu S R, Chudgar H S. A novel approach towards utilizing Dempster Shafer fusion theory to enhance WiFi positioning system accuracy. Pervasive and Mobile Computing, **37**(1): 115–123(2017)

[7] Jiao J, Deng Z, Xu L, et al. A hybrid of smartphone camera and basestation wide-area indoor positioning method. KSII Transactions on Internet and Information Systems (TIIS), **10**(2): 723–743(2016)

[8] Jin M, Liu S, Schiavon S, et al. Automated mobile sensing: Towards high-granularity agile indoor environmental quality monitoring. Building and Environment, **127**(1): 268-276(2018)

[9] Heath M, Sarkar S, Sanocki T, et al. Comparison of edge detectors: a methodology and initial study. Computer vision and image understanding, **69**(1): 38–54(1998)

[10] FischlerM A, BollesR C.Random sample consensus:aparadigm for model fitting with application to image analysis and automated cartography.Communication Association Machine **24**(6) :381-395(1981).

Research on Air pollution Spatiotemporal Forecast Model Based on LSTM

JingWei Yu, Hongyang Yu

***Abstract*—At present, the increasingly serious air pollution in various cities of China has made people pay more attention to the air quality index(hereinafter referred to as AQI) of their living areas. To face this situation, it is of great significance to predict air pollution in heavily polluted areas. In this paper, based on the time series model of LSTM, a spatiotemporal prediction model of PM2.5 concentration in Mianyang, Sichuan Province is established. The model fully considers the temporal variability and spatial distribution characteristics of PM2.5 concentration. The spatial correlation of air quality at different locations is based on Air quality status of other nearby monitoring stations, including AQI and meteorological data to predict the air quality of a monitoring station. The experimental results show that the method has good prediction accuracy that the fitting degree with the actual measured data reaches more than 0.7 which can be applied to the modeling and prediction of the spatial and temporal distribution of regional PM2.5 concentration.**

***Index Terms*—LSTM,PM2.5,Deep Learning,AQI**

I. Introduction

Particulate pollutants are one of the most serious pollutants affecting human health. PM10 is usually defined as inhalable particulate matter with an aerodynamic diameter less than 10um. PM10 can be inhaled by the human respiratory tract and penetrate deep into the human lung, stimulating the pulmonary capillaries and causing difficulty in breathing; while $PM_{2.5}$ is defined as aerodynamic Fine particles with a diameter of less than $2.5\mu m$ can break through the lung air-blood barrier and enter the human blood system, and because the complex structure aggregates composed of them have a larger surface area than the simple structure aggregates composed of large particles, they are easier to adsorb Some heavy metals and organic substances that are harmful to human health have higher toxicity. It usually affects the respiratory, circulatory and central nervous systems, body metabolism and immunity, genitourinary system, blood system, digestive system, and skin to varying degrees [**?**].

In my country, the problem of air pollution is becoming more and more serious. Since the reform and opening up, my country's economy has been rising steadily, and the modernization process has been advancing rapidly. However, the development method in the early stage of reform and opening up has brought important losses and problems to my country's ecological environment. Extreme air pollution events occur frequently throughout my country. In 2020, among the 337 cities in the country, 202 cities met the environmental air quality standard, accounting for 59.9% of all cities, up 13.3 percentage points from 2019; 135 cities exceeded the standard, accounting for 40.1%, down 13.3 percentage points from 2019 percent. A total of 345 days of severe pollution occurred in 337 cities, and the days with $PM_{2.5}$, $PM_{1}0$ and O3 as the primary pollution accounted for 77.7%, 22.0% and 1.5% of the days with severe and above pollution, respectively. At the beginning of 2020, under the influence of unfavorable meteorological conditions, large-scale $PM_{2.5}$ pollution occurred across the country. Heavy $PM_{2.5}$ hourly pollution first appeared in central Liaoning and Guanzhong areas, and then spread widely. Among the 337 cities above the prefecture level, 10 cities were severely polluted, and the peak $PM_{2.5}$ hourly concentration reached $365\mu g/m^3$. At the beginning of May 2020, Beijing and surrounding areas were affected by dry and hot weather. $PM_{2.5}$ concentrations reached moderate to severe pollution levels. Factors such as the lack of accurate input data (such as emission sources and emissions) for numerical forecasting models may cause these models to differ in their predicted results. This leads to the unsatisfactory performance of numerical prediction models in real-time prediction. On the other hand, with the continuous improvement of the development level of sensors, the availability of historical data collected by different sensors has been greatly improved. Observation-based statistical forecasting techniques are therefore another widely used approach, in which statistical models link several explanatory variables to predict $PM_{2.5}$ concentrations as output [5]. Researchers from the Chinese Academy of Sciences proposed a spatiotemporal convolutional long short-term memory neural network extension model (C-LSTME) to extract high-level spatiotemporal features through the combination of convolutional neural network and long short-term memory neural network (LSTM-NN). And integrate meteorological data and aerosol data to improve model prediction performance. The results of the model achieved an accuracy of 87.6% compared to the actual observed ranking [1]. Some researchers propose an optimal hybrid model that combines the advantages of quadratic decomposition (SD), artificial intelligence methods and optimization algorithms, select wavelet decomposition as the main decomposition technology, and then use LSTM to simplify the prediction, and finally use LSSVM to obtain the final result, This model can fully capture AQI features and has a high accuracy rate [2]. To sum up, it can be concluded that the research on air pollution forecasting using artificial neural network or deep learning and other artificial intelligence methods has blossomed everywhere, and most of

University of Electronic Science and Technology of China, Chengdu,China, e-mail: yujingweiop@gmail.com.

the researches have very gratifying test forecasting effects on their test data sets.

II. Data and methods

A. Study area and available data

Mianyang, a prefecture-level city in Sichuan Province, is located in the northwest of the Sichuan Basin, in the middle and upper reaches of the Fujiang River. Between $0°42' - 33°03'$ north latitude and $103°45' - 105°43'$ east longitude.In 2021, the number of days with good ambient air quality in the urban area will be 80.0%, and the comprehensive air quality index will be 4.78. Among them, $PM_{2.5}$ is 55.8 micrograms per cubic meter, PM10 is 87.3 micrograms, and ozone is 107.3 micrograms. The data is released by the Environmental Protection Administration of China, including 29 monitoring stations in Mianyang City, with a total of 352,423 pieces of air quality data and meteorological data [2]. The data collection interval is one hour. The data collection period was from June 5, 2019 to January 31, 2020.

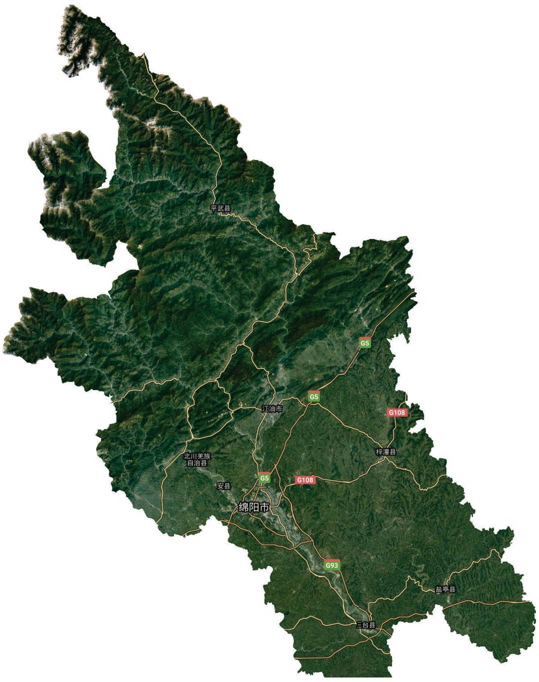


Fig. 1. the location of mianyang city

B. Spatial dependence modelling

Since the air quality of the whole city cannot be determined by a single monitoring station or several monitoring stations, the air quality of different areas in a city is usually different. In view of this situation, the prediction of the air quality of the whole city through monitoring station data is not accurate enough. It is feasible to make separate forecasts for monitoring stations to achieve forecasts for a small area [3]. In this paper, cluster analysis is used to find out the neighboring monitoring stations that have a greater impact on the A monitoring station, as spatial feature extraction. The purpose of cluster analysis is to divide a data set into different categories (clusters), and the desired goal is to maintain similar characteristics of objects in a category. Exploring the internal structure of data is the most widely used direction of cluster analysis. For example, it can be applied to customer purchase data to find potential customer purchase preference categories (also known as clusters or clusters). The number of this category can be manually determined by the experimenter in advance, or it can be calculated by the algorithm itself, which is determined by the specific application and the clustering method used. Since the goal of clustering is not to simply judge that a single variable belongs to a specific category, it is necessary to calculate the similarity between observations in each cluster from an overall perspective, find all observations with similar characteristics and classify them into the same category. Chebyshev Distance is a measure in vector space, which defines the distance between two points in space coordinates as the maximum value of the absolute value of the difference between the coordinates of their coordinates. Chebyshev distance is the number of moves the king takes to move from one square to another in a chess board:

$$dict_{cd} = \lim_{t}(\sum_{k=1}^{m} |x_{ik} - x_{jk}|^{t})^{\frac{1}{t}} \tag{1}$$

Minkowski Distance is a measure of Euclidean space, a definition of a set of distances, and is regarded as a generalization of Euclidean distance and Manhattan distance.

$$dict_{mind} = p\sqrt{\sum_{k=1}^{m} |x_{ik} - x_{jk}|^{p}} \tag{2}$$

III. Temporal dependency modelling

Long short-term memory neural network (Long short-term memory, LSTM) is a special RNN that is improved to solve the problem of gradient disappearance and gradient explosion in RNN in the long sequence training process. Compared with ordinary RNN, LSTM is a good solution to the problem that RNN cannot handle long-term dependencies. After being introduced by Hochreiter and Schmidhuber (1997) [4], it is still a widely used neural network.The repetitive work module of LSTM is shown below.

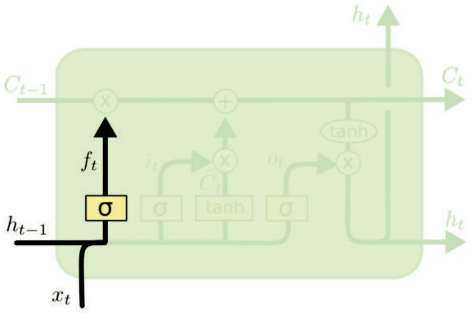


Fig. 2. The localsation of mianyang city

The basic unit of LSTM is the cell stateC_t, and the current C_t is jointly determined by the forgetting gate and the input gate. Next, the functions and principles of each part will be introduced in detail. In the following, for the convenience of explanation and use, define the forgetting gatef_t, the input gatei_t, and the output gateh_t.

1) Forgotten Gate

The forget gate f_t is the first step in the LSTM calculation. The forget gate receives the cell state C_{t-1} output by the previous unit and decides how much information is retained and how much is discarded. The signal of the forget gate f_t is obtained by performing $Sigmoid$ in the $Sigmoid$ computing unit by the current input X_t and the calculation result h_{t-1} output by the previous unit. Finally, f_t and the cell state C_{t-1} are multiplied to determine the proportion of information retention. The formula for forgetting gatef_t is:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \tag{3}$$

Among them, f_t is recorded as the output of the forget gate, W_f is recorded as the weight matrix, and b_t is recorded as the bias. Then the resulting f_t is between $0 \sim 1$, 1 means to keep all the information, on the contrary 0 discards all the information. The specific operation flow of the forget gate is shown in the Fig.3.

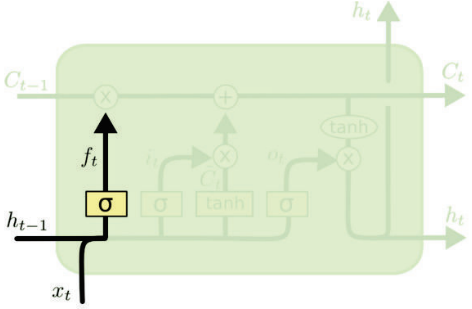


Fig. 3. The flow of forgotten gate

2) input gate

The input gate of LSTMi_t mainly determines the information composition of the current cell stateC_t. There are three operation steps, the specific steps are as follows: The $Sigmoid$ operation is performed on the h_{t-1} and X_t inputted by the current cell state to obtain the signali_t of the new information retention ratio, which is similar to the operation of the forget gate.
Perform tanh operation on h_{t-1} and X_t of the current cell state to obtain the candidate cell state to be input $\tilde{C}_t$;
Multiply the result obtained in step a with the candidate input information $\tilde{C}_t$ in step b to obtain the cell stateC_t that needs to be retained to the current LSTM this time.

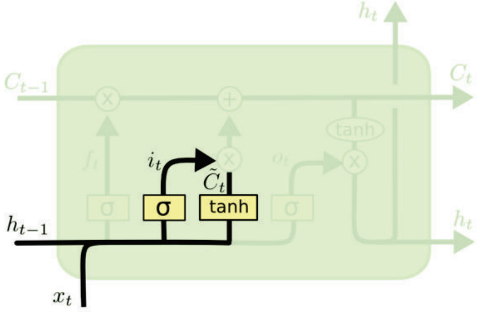


Fig. 4. The flow of input gate

3) Cell state update

The information of the cell state C_t of the LSTM will be updated in this step. First, the signal f_t of the forget gate is used to determine how much information is retained in the cell state C_{t-1} of the previous step, so that the cell state of the LSTM forgets part of the information, and then the current cell state $tildeC_t$ is added. At this point, the cell state C_t of the LSTM is updated. The specific process of cell state update is shown in the Fig.5:

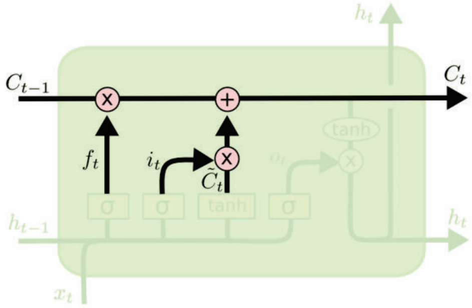


Fig. 5. The flow of cell state update

The calculation formula of cell state update is:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{4}$$

4) output gate

Finally, calculate the output information h_t of the LSTM model in this cell state. First, a sigmoid layer is used to decide which parts of the cell state to output. Similar to the result of the forget gate, the sigmoid operation is performed with f_{t-1} and X_t to obtain the output gate ratio O_t. Similar to the operation of the input gate, LSTM obtains the pre-output information C_t after the cell state C_t of this step undergoes a tanh operation. Finally, it is multiplied with the output gate signal o_t, and after screening, the output information h_t of the LSTM in the current cell state is obtained. This output is the input to the iteration of the next cell state of the LSTM. The specific operation flow of the calculation output is shown in the Fig.6.
The calculation formula of LSTM is sorted as follows:

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{5}$$

$$h_t = o_t * tanh(C_t) \tag{6}$$

The temporal prediction part uses LSTM neural network to predict $PM_(2.5)$ concentration. The LSTM neural net-

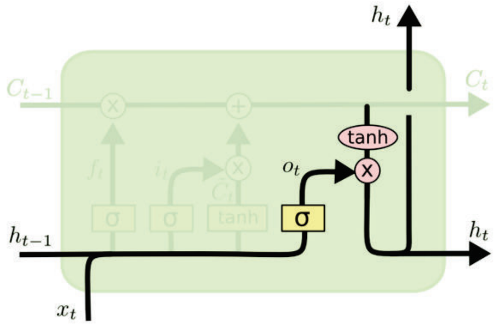


Fig. 6. The flow of output gate

work uses four different types of data to simulate the trend of $PM_(2.5)$ air quality at a monitoring station: 1) The $PM_(2.5)$ value of this monitoring station in the past h hours, the $PM_(2.5)$ value can be regarded as equal to the AQI value; 2) Meteorological data of the tc monitoring station at the current time point (such as weather conditions, sunny/cloudy/cloudy/foggy, humidity conditions, wind speed, wind speed direction, etc.); 3) the duration of the time step of the interval; 4) with our Meteorological data for the same time interval as the time interval to be forecasted.

Obviously, the current air state and meteorological state have different degrees of influence on different time intervals in the future. Therefore, as shown in the Fig.7, we formulate different training sets for the input (indicated by the dotted matrix) and the air quality t(c+1) at different time intervals, which are respectively used to train different models corresponding to different time intervals. Each blue dashed arrow in the Fig.7 represents a temporal predictor. We divide the next 8 hours into two time intervals $0 \sim 1$hour$1 \sim 8$hour. This paper uses a model to predict the AQI value for each time interval, the first three types of data used are the same in different time forecasts, and the input data of the forecasters for different time intervals is different from the meteorological data.

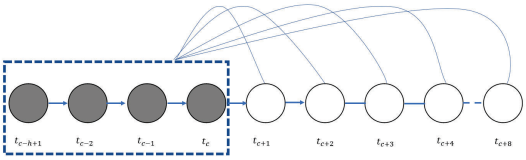


Fig. 7. Schematic diagram of the time predictor

Hyperparameter settings 1. Batch size (Batchsize) According to the GPU performance and memory capacity used, set the batch size to 32, how many batches to divide the training set into, and the completion of all batches of training is regarded as the completion of one round (epoch) of training. 2. Training Epoch (Epoch) During the model training optimization process, when the training error does not drop significantly, the training stops. The model training rounds set in this study are 100 rounds. 3. Loss Function In this study, the mean square error loss function (MAE) is selected, and the smaller the error, the better the prediction performance of the model. 4. Optimizer In this study, Adam optimizer suitable for MAE loss function is used to optimize the training of the model. 5. Learning Rate In this study, the learning rate is selected as 0.0001. During the training process, the learning rate will be adjusted automatically according to the training progress.

IV. Conclusion

In this paper, to predict the air quality of a monitoring station, the real current air quality can be obtained from future records, and the data six hours before the current time is selected as the prediction step. Predict the air quality in the next hour. The baseline model used in this paper is the autoregressive moving average (ARMA) as the baseline. ARMA is a well-known time series data prediction model. ARMA only predicts the air quality of the site based on the air quality index of the site. This paper uses RMSE and MAE as the evaluation criteria of the model. The method used in this paper can be simply referred to as the FA model. The time forecast in this paper is 30.41 for RMSE and 23.93 for MAE for 1-4 hours. RMSE for 1-8 hours is 39.97 and MAE is 28.42. The time prediction results of the baseline ARMA were $1 \sim 4$ hours RMSE 33.25, MAE 27.78. $1 \sim 8$ hours RMSE 48.64, MAE 33.12. Our FA model outperforms the baseline ARMA in both short- and long-term predictions.

TABLE I
ARMA'S RESULT

Prediction	ARMA	
	RMSE	MAE
+4h	33.25	27.78
+8h	48.64	33.12

TABLE II
FA'S RESULT

Prediction	FA	
	RMSE	MAE
+4h	30.41	23.93
+8h	39.97	28.42

The FA model shows better performance than the baseline. The main reason is that meteorological data is considered, and the second feature is extracted from geographic information. This makes the FA model outperform the baseline not only in the short term (within 4 hours) but also in the medium and long term (within 8 hours). However, there are also disadvantages. First, the mutation factor is not considered, and the ability to predict the mutation weather is relatively weak. The second point is that the handling of missing values in time series is not adequate and needs to be improved.

References

[1] Wen C, Liu S, Yao X, et al. A novel spatiotemporal convolutional long short-term neural network for air pollution prediction[J]. Science of the total environment, 2019, 654: 1091-1099.

[2] Zhou Q, Jiang H, Wang J, et al. A hybrid model for PM2. 5 forecasting based on ensemble empirical mode decomposition and a general regression neural network[J]. Science of the Total Environment, 2014, 496: 264-274.

[3] Zheng Y, Yi X, Li M, et al. Forecasting fine-grained air quality based on big data[C]//Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. 2015: 2267-2276.

[4] Graves A. Long short-term memory[J]. Supervised sequence labelling with recurrent neural networks, 2012: 37-45.

[5] Ma J, Li Z, Cheng J C P, et al. Air quality prediction at new stations using spatially transferred bi-directional long short-term memory network[J]. Science of The Total Environment, 2020, 705: 135771.

A Soft Error Rates (SER) Evaluation Method of Combinational Logic Circuit Based on Linear Energy Transfers

Man Li, Wanting Zhou, and Lei Li

Abstract—Communication stability is the primary concern of communication satellites. Communication satellites are easily affected by particle radiation to generate single event effects (SEE), which leads to soft errors (SE) of combinational logic circuit. The existing research on soft error rates (SER) of combined logic circuit is mostly based on the assumption that the logic gates being bombarded have the same pulse width. However, in the actual radiation environment, the pulse widths of the logic gates being bombarded are different due to different linear energy transfers (LET). In order to improve the accuracy of SER evaluation model, this paper proposes a soft error rates evaluation method based on LET. In this paper, the authors analyze the influence of LET on the pulse width of combinational logic and establish the pulse width model based on LET. Based on this model, the error rate of test circuit ISCAS'85 is calculated. Experimental results show that this model can be used for SER evaluation.

Keywords—Communication satellite, pulse width, soft error rates, LET.

I. Introduction

THE reduction in process size will lead to the increase of communication satellites sensitivity to radiation particles [1], [2]. The injection of particles into the node of an electronic device will result in the accumulation of charge at the node and the generation of single event transients (SET). SETs propagating in combinational logic circuit may be stored by timing element, resulting in circuit output error and affecting the normal operation of the circuit. This error is called soft errors and soft error rates (SER) represent the vulnerability of a circuit to SETs [3]. It is of great significance to establish an accurate SER evaluation model for the anti-radiation research of the combinational logic circuit.

Many scholars have done a lot of research on SER evaluation [4]-[8]. Baojun Liu et al. proposed a Monte Carlo model to analyze the reliability of transient pulse on combinational logic circuits [9]. Three masking effects, logical masking, electrical masking, and latch window masking were considered in this model. Georgios Ioannis Paliaroutis et al. established a soft error evaluation model based on layout information and considered the influence of temperature on the pulse widths of SETs [10]. Anglada, M et al. introduced an innovative way which combined signal probabilities with technology characterization to calculate the SER [11]. Farjaminezhad, R et al. proposed a shape prediction method for transient faults based on recursive neural network (RNN) [12]. This method could be used to estimate the influence of single or multiple transient faults propagating in a combined circuit. Cai, S et al. presented a methodology of multi-transient fault simulation based on probability distribution, in which the logical output results were subject to Bernoulli distribution [13].

The probability of a soft error caused by SETs depends on the SET pulse widths [14]. A larger SET pulse width is more easily propagated in the combinational logic circuit and stored by triggers [15]. The existing research on soft error rates (SER) of combined logic circuit is mostly based on the assumption that the logic gates being bombarded have the same pulse width. Whereas, in the actual radiation environment, the pulse widths of the logic gates being bombarded are different due to different linear energy transfers (LET). In this paper, a soft error rates evaluation model based on LET was proposed. The authors analyzed the effect of LET on the SET pulse width, studied the modeling of SET pulse width, and finally obtained the soft error rates of combinational logic circuits by circuit level simulation.

The structure of this paper is as the following: Section II introduces the affection of LET on SET pulse widths, Section III proposes the SER evaluation model based on LET. Section IV displays the effects of LET on the SER and verifies the validity of the model.

II. The Affection of LET on SET Pulse Widths

In this paper, the effect of LET on SET pulse widths was analyzed with the basic gate (inverters, NOR gates, and NAND gates) as the research object. This paper took an inverter as an example to present the simulation process. A device-circuit level hybrid simulation model of the basic gates was built on TCAD platform, which had been calibrated with reference to the 40 nm process, as shown in Fig. 1. The NMOS bombarded by particles was the device model, and the rest was the SPICE model. Table. I shows the structure and doping parameters of the NMOS.

The heavy ion radiation environment was simulated on the established 3D model. The simulation location was set as the drain center of the NMOS, the direction was vertical incident, the incident particle was heavy ion, and the value of LET ranged from 0.01 pc/μm to 0.05 pc/μm. Fig. 2 reflects the change of SET voltage pulse of inverters. It can be seen

Man Li is with Research Institute of Electronic Science and Technology, University of Electronic Science and Technology of China, Chengdu 611731, China. (e-mail: manlucky2022@163.com).

Wanting zhou is with Research Institute of Electronic Science and Technology, University of Electronic Science and Technology of China, Chengdu 611731, China. (e-mail: zhouwt@uestc.edu.cn.).

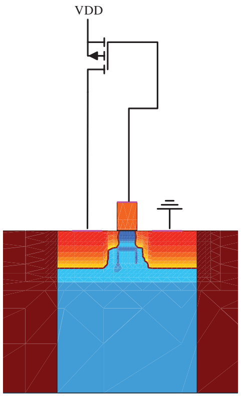


Fig. 1. The structure of inverters

TABLE I
THE STRUCTURE AND DOPING PARAMETERS OF THE NMOS

Parameter	Value
The length of the channel(nm)	55
Gate oxide layer thickness(nm)	1.25
Source and drain doping concentration(cm-3)	1.00e21
VT doping concentration(cm-3)	1.29e19
Substrate doping concentration(cm-3)	6.00e16

from the figure that the width of SET voltage pulse window increases with the increase of LET. The reason was as follows: With the increase of LET, the concentration of electron hole pair generated by particle incident increased, and the drain needed a longer diffusion time to recover to the initial state. SET pulse width was defined as FWHR (the full width of half rail). The value of FHWR was calculated according to the following expression.

$$FWHR = T_1 - T_2 \tag{1}$$

where T_1 is the time to recover to half of voltage amplitude, T_2 is the time to drop to half of voltage amplitude.

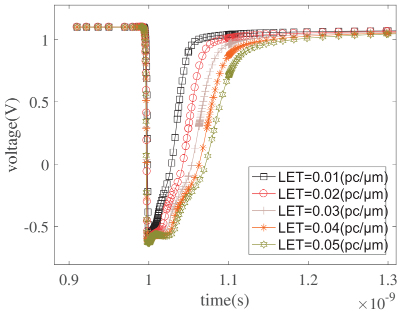


Fig. 2. The change of SET voltage pulse of inverters

According to the above expression, the SET pulse width of the basic gates was obtained. As shown in Fig. 3, similar to the inverters, the NOR gates, and NAND gates also conform to the phenomenon that the pulse width increases as the LET increases. It can also be seen that the SET pulse widths of NAND gates and NOR gates are larger than that of inverters.

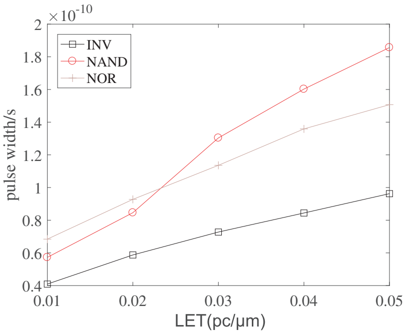


Fig. 3. SET voltage pulse of base gates

III. A SER Evaluation Model Based on LET

Based on the above research on the influence of LET on SET pulse widths, this section proposed a LET-based soft error rates evaluation model, which took into account logical masking, electrical masking, latching masking and the random injection of SET pulse. SET pulse widths of the basic gates were modeled by the Polyfit function in the MATLAB. The pulse widths of the basic gates were obtained as follows:

$$y_1 = 9.34e{-}10 \cdot x + 4.61e{-}11 \tag{2}$$

$$y_2 = -6.74e{-}10 \cdot x^2 + 1.80e{-}9 \cdot x + 5.95e{-}11 \tag{3}$$

$$y_3 = \begin{cases} -2.07e{-}9 \cdot x^2 + 3.19e{-}9 \cdot x + 3.65e{-}11 & x < 0.4 \\ 1.01e{-}9 \quad 0.4 \leq x \leq 0.6 \end{cases} \tag{4}$$

where y_1 is the pulse width of inverters, y_2 is the pulse width of NOR gates, y_3 is the pulse width of NAND gates, x is the value of LET.

Fig. 4 shows the structure diagram of the soft error rates evaluation model, including the following parts.

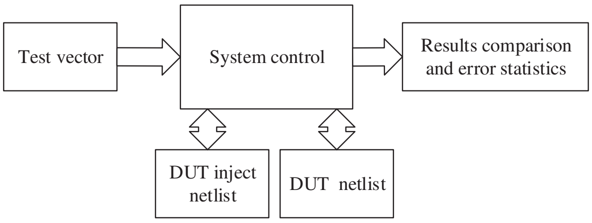


Fig. 4. The structure diagram of the proposed model

1) System control. This part mainly included test vector address generation, SET pulse width generation, injection time control and injection position control. The pseudocode is shown in Table. II.
2) DUT inject netlist. DUT inject netlist was the netlist file for the circuit under test. The DC synthesis tool was used to synthesize the circuit under test into gate level

netlists containing only inverters, NAND gates and NOR gates. The pulse injection model, derived from the Pulse model, was inserted into each of the base gates in the netlist.

3) DUT netlist. Unlike DUT inject netlist, this netlist did not insert the pulse injection model and was used to compare with the output of DUT inject netlists.
4) Test vector. Traversal of all inputs would greatly increase the simulation time. In order to save time, random vector was used as the input of the netlist.
5) Results comparison and error statistics. Compare the output results of DUT Inject netlists and DUT Netlists by XOR operation. If the value was 0, there was no error. Otherwise, the error count was increased by 1.

TABLE II
THE PSEUDOCODE OF SYSTEM CONTROL

```
Algorithm : system ctrol
Setting the number of test vector
Setting the injection time
Setting the injection location
Setting the number of LET
For netlist in the list of netlist
  For LET in the list of LET
  set the pulse-width-inv as the pulse-width of inverters by (2)
  set the pulse-width-nor as the pulse-width of NOR gates by (3)
  set the pulse-width-nand as the pulse-width of NAND gates by (4)
    For injection location in the list of injection location
      For a random input vector in the input vector
        For injection time in the list of injection time
        Simulate the target netlist
        output-inject <- the output of DUT inject netlist
        output-unject <- the output of DUT netlist
        if (output-unject!=output-inject)
          counter-error ++
          END
        END
      END
    END
  END
END
```

IV. Analysis of Results

A. Analysis of SET pulse width model

The SET pulse width model was established by fitting the SET pulse widths simulated by TCAD. The value of LET used for SET pulse width modeling ranged from 0.01 pc/μm to 0.6 pc/μm. MSE and regression coefficient (R^2) were taken as metrics of the model accuracy.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{5}$$

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{6}$$

where n is the number of data; y_i is the factual data and $\hat{y}_i$ is the predicted data.

The closer regression coefficient is to 1, the smaller MSE is, indicating that the fitting effect is better. Fig. 5 shows the comparison between TCAD simulation results and this model. From Table. III and Fig. 5, it could be found that both were quite close and could be used for subsequent research.

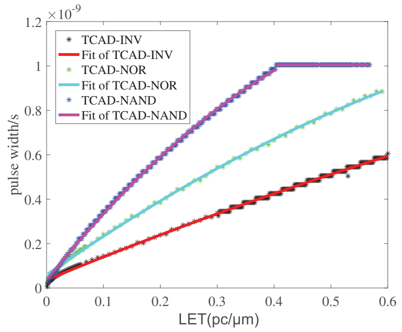


Fig. 5. The comparison between TCAD and this model

TABLE III
MODEL INDICATORS

base gates	R^2	MSE
inverters	0.99715	8.39e-23
NAND gates	0.99936	3.02e-20
NOR gates	0.99924	5.79e-23

B. Analysis of SER model

Using the LET-based SER evaluation model proposed above, this paper analyzed the soft error rates under different LET of part of circuits belonging to the ISCAS'85 benchmark set. The SMIC 40 nm process library was designated as the comprehensive target library in this paper. The number of test vectors was 150 and the value of LET ranged from 0.1 pc/μm to 0.5 pc/μm. The operating environment of this model was as follows: CPU was Intel Core i5-8500, GPU was NVIDIA GeForce GTX 1060 5GB.

Fig. 6 presents the soft error rates of the ISCAS'85 Benchmark circuits. As LET increased, the probability of circuit error increased. This phenomenon was related to the SET pulse widths. As LET increased, SET pulse widths increased, and the probability of pulse propagating in combination logic and being locked by timing unit also increased. The probability of a SET pulse being stored increased, and so did the probability of an output error.

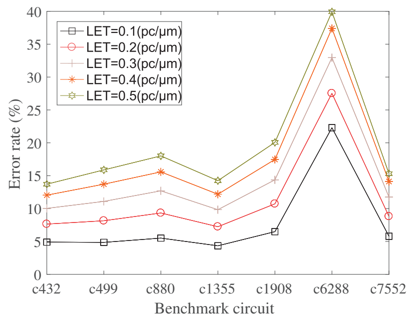


Fig. 6. The error rates of the ISCAS'85 benchmark circuits

The validity of the model was demonstrated by comparing with [9]. The circuit reliability analysis data in [9] when pulse width was 200 ps was compared with the error rates in this model when LET was 0.1pc/μm. Set the LET value of this model to 0.1 pc/μm, which meant that the pulse width of inverters was 140 ps, the pulse width of NOR gates was 233 ps, and the pulse width of NAND gates was 335 ps. Fig. 7 shows the data comparison between this model and [9]. As can be seen from Fig. 7, the trend of data in this model is almost the same as that in [9]. It should be noted that, in general, various algorithms to study the soft error rates would carry out different simplified processing, and this model adopted the 40 nm process, while the comparison model adopted the 65 nm process. Different process technologies would also cause deviations in the results.

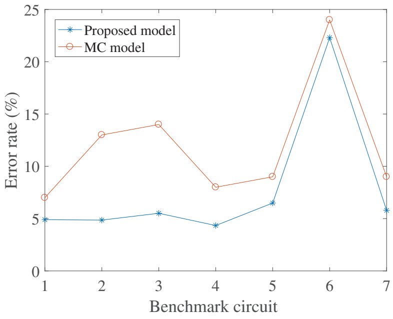


Fig. 7. The error rates of the proposed model and [9]

The model was set as two cases of the same and different pulse widths of the basic gates. By comparing the error rates in these two cases, the research significance of the model was further proved. In the first case, the basic gate used the same pulse width of 140 ps; in the second case, the authors set the LET value as 0.1 pc/μm, and the pulse widths corresponding to the basic gates were 140 ps, 233 ps and 335 ps respectively. As you can see from Fig. 8, in the first case, the soft error rates may be underestimated, resulting in an inaccurate assessment of the soft error rates. Similarly, the soft error rates may be exaggerated when the same pulse width was 335 ps.

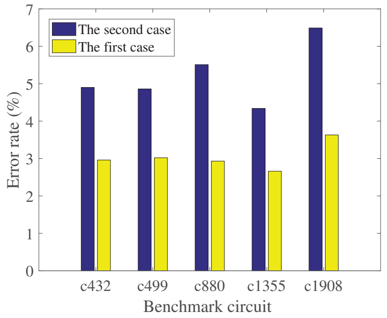


Fig. 8. The error rates in case one and the error rates in case two

V. Conclusion

In this work, A SER evaluation method of combinational logic circuit based on LET was proposed to predict SER under different LETs. Compared with the model using the same pulse width, the SER prediction was more accurate. If the influence of LET on the pulse widths of the basic gates was not considered, the soft error rates assessment would be inaccurate. This model can be applied to the study of satellite radiation resistance.

Acknowledgment

This work was supported by Sichuan Science and Technology Program (No.2021YJ0082), the National Defense Science and Technology Key Laboratory Fund (No.6142103190310) and the National Natural Science Foundation of China NSAF Joint Fund (Grant No.U1630133).

References

[1] Lwin, Ne Kyaw Zwa and Sivaramakrishnan, “Single-Event-Transient Resilient Memory for DSP in Space Applications,” 2018 IEEE 23rd International Conference on Digital Signal Processing (DSP), pp. 1-5 Nov. 2018.

[2] T. Li, H. Liu and H. Yang, "Design and Characterization of SEU Hardened Circuits for SRAM-Based FPGA," in IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 27, no. 6, pp. 1276-1283, June 2019, doi: 10.1109/TVLSI.2019.2892838.

[3] M. R. Rohanipoor, B. Ghavami and M. Raji, "Improving Combinational Circuit Reliability Against Multiple Event Transients via a Partition and Restructuring Approach," in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 39, no. 5, pp. 1059-1072, May 2020, doi: 10.1109/TCAD.2019.2907922.

[4] T. Uemura et al., "Investigation of alpha-induced single event transient (SET) in 10 nm FinFET logic circuit," 2018 IEEE International Reliability Physics Symposium (IRPS), 2018, pp. P-SE.1-1-P-SE.1-4, doi: 10.1109/IRPS.2018.8353689.

[5] N. Pande, S. Kumar, L. R. Everson and C. H. Kim, "Understanding the Key Parameter Dependences Influencing the Soft-Error Susceptibility of Standard Combinational Logic," in IEEE Transactions on Nuclear Science, vol. 67, no. 1, pp. 116-125, Jan. 2020, doi: 10.1109/TNS.2019.2959922.

[6] S. A. Olowogemo, A. Yiwere, B. -T. Lin, H. Qiu, W. H. Robinson and D. B. Limbrick, "Electrical Masking Improvement with Standard Logic Cell Synthesis Using 45 nm Technology Node," 2020 IEEE 63rd International Midwest Symposium on Circuits and Systems (MWSCAS), 2020, pp. 619-622, doi: 10.1109/MWSCAS48704.2020.9184651.

[7] M. R. Rohanipoor, B. Ghavami and M. Raji, "Improving Combinational Circuit Reliability Against Multiple Event Transients via a Partition and Restructuring Approach," in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 39, no. 5, pp. 1059-1072, May 2020, doi: 10.1109/TCAD.2019.2907922.

[8] Z. Zhang et al., "Extrapolation Method of On-Orbit Soft Error Rates of EDAC SRAM Devices From Accelerator-Based Tests," in IEEE Transactions on Nuclear Science, vol. 65, no. 11, pp. 2802-2807, Nov. 2018, doi: 10.1109/TNS.2018.2875051.

[9] B. Liu and L. Cai, "Monte Carlo Reliability Model for Single-Event Transient on Combinational Circuits," in IEEE Transactions on Nuclear Science, vol. 64, no. 12, pp. 2933-2937, Dec. 2017, doi: 10.1109/TNS.2017.2772267.

[10] G. Ioannis Paliaroutis, P. Tsoumanis, N. Evmorfopoulos, G. Dimitriou and G. I. Stamoulis, "A Placement-Aware Soft Error Rate Estimation of Combinational Circuits for Multiple Transient Faults in CMOS Technology," 2018 IEEE International Symposium on Defect and Fault Tolerance in VLSI and Nanotechnology Systems (DFT), 2018, pp. 1-6, doi: 10.1109/DFT.2018.8602855.

[11] M. Anglada, R. Canal, J. L. Aragón and A. González, "Fast and Accurate SER Estimation for Large Combinational Blocks in Early Stages of the Design," in IEEE Transactions on Sustainable Computing, vol. 6, no. 3, pp. 427-440, 1 July-Sept. 2021, doi: 10.1109/TSUSC.2018.2886640.

[12] Farjaminezhad R , Safari S , Moghadam A . "Recurrent Neural Networks Models for Analyzing Single and Multiple Transient Faults in Combinational Circuits." Microelectronics Journal, 2021(7553):104993.

[13] Cai S , He B and Wang W. "Soft Error Reliability Evaluation of Nanoscale Logic Circuits in the Presence of Multiple Transient Faults". Journal of Electronic Testing, vol. 36, 2020.

[14] B. Narasimham et al., "On-Chip Characterization of Single-Event Transient Pulsewidths," in IEEE Transactions on Device and Materials Reliability, vol. 6, no. 4, pp. 542-549, Dec. 2006, doi: 10.1109/TDMR.2006.885589

[15] H. Pahlevanzadeh and Q. Yu, "Systematic analyses for latching probability of single-event transients," Fifteenth International Symposium on Quality Electronic Design, 2014, pp. 442-449, doi: 10.1109/ISQED.2014.6783358.

PLACE
PHOTO
HERE

Michael Shell Biography text here.

John Doe Biography text here.

Jane Doe Biography text here.

Sea-Land Segmentation Method Based on the Transformer with Enhanced Edge Supervision

Lianzhong Zhang, Chao Huang

Abstract**—**Sea-land segmentation is a basic step in many tasks such as sea surface monitoring and ship detection. The existing sea-land segmentation algorithms have poor segmentation accuracy, and the parameter adjustments are cumbersome and difficult to meet actual needs. Also, the current sea-land segmentation adopts traditional deep learning models that use Convolutional Neural Networks (CNN). At present, the transformer architecture has achieved great success in the field of natural images, but its application in the field of radar images is less studied. Therefore, this paper proposes a sea-land segmentation method based on the transformer architecture to strengthen edge supervision. It uses a self-attention mechanism with a gating strategy to better learn relative position bias. Meanwhile, an additional edge supervision branch is introduced. The decoder stage allows the feature information of the two branches to interact, thereby improving the edge precision of the sea-land segmentation. Based on the Gaofen-3 satellite image dataset, the experimental results show that the method proposed in this paper can effectively improve the accuracy of sea-land segmentation, especially the accuracy of sea-land edges. The mean IoU (Intersection over Union), edge precision, overall precision, and F1 scores respectively reach 96.36%, 84.54%, 99.74%, and 98.05%, which are superior to those of the mainstream segmentation models and have high practical application values.

Keywords**—**Sea-land segmentation, transformer, deep learning, SAR.

I. Introduction

SEA-LAND segmentation is necessary for a variety of tasks such as coastline extraction and monitoring, marine oil spills, pollution monitoring, tide detection, and ship detection [1]-[5]. However, as for the synthetic aperture radar (SAR) image, its imaging methods, geometric characteristics, and radiation characteristics are different from natural images [6], [7]. The light spots and complex sea clutter caused by the mutual interference of multiple objects can easily lead to images of poor quality, and the edges of the sea-land are difficult to distinguish [8]-[10]. Therefore, the sea-land segmentation of SAR images is challenging.

At present, the methods of sea-land segmentation are mainly divided into two categories. The first is traditional SAR image segmentation methods, including the segmentation methods based on hybrid models and the segmentation methods based on thresholds, etc. [11]-[14]. Such methods are susceptible to noise interference, and their parameters need to be adjusted appropriately, making the methods less robust. The second is deep learning methods that are widely used in the field of image segmentation because of their powerful image feature extraction capabilities [15]-[17]. At present, most sea-land segmentation methods are improved and optimized referring to natural image segmentation models, such as UNet [18], SegNet [19], and Deeplab [20] series. These models are all built based on the CNN architecture and have achieved good results in optical images. Although CNN has good local feature extraction capabilities, it extracts abstract high-level features through continuously stacked convolution kernels, and it can also have a theoretical receptive field covering the entire image. A plethora of studies has proved that the actual receptive field of CNN is smaller than the theoretical receptive field [21], which is not conducive to making full use of context information to obtain more accurate segmentation features. To overcome the inherent deviations of the CNN architecture, the transformer architecture emerges, which has achieved great success in the field of natural images [22], [23]. It calculates global feature information with the self-attention mechanism and can better handle long-distance features. Therefore, this paper proposes a sea-land segmentation method based on the transformer architecture. The main contributions of this work are summarized as follows:

1) The transformer architecture network is applied to the sea-land segmentation of SAR images;

2) Aiming at the problem that the position bias is difficult to learn under a small SAR image dataset (the SAR image generally has less data), this paper proposes a self-attention mechanism with a gating strategy that can better learn the position bias and improve the segmentation accuracy;

3) To reduce the segmentation error caused by the scattering features at the edge of the sea-land in the SAR image, this paper designs additional edge supervision. The features extracted by the global CNN branch are merged in the shallow features of the encoder, and the learned segmentation features are combined in the decoder stage. The edge features interact with each other and help to

Lianzhong Zhang is with Research Institute of Electronic Science and Technology, University of Electronic Science and Technology of China (UESTC), China (corresponding author, e-mail: zlz0615@foxmail.com).

Chao Huang is with Computer Science and Engineering College, University of Electronic Science and Technology of China (UESTC), China (e-mail: chaohuang98@gmail.com).

predict the sea-land mask and the sea-land edge at the same time, contributing to more accurate sea-land segmentation.

II. Our Method

Fig. 1 illustrates the main structure of the network proposed in this paper, which consists of four parts: encoder, decoder, additional edge supervision branch, and global CNN branch. The swin-transformer [24] is used as the backbone, which is the state-of-the-art (SOTA) in the field of computer vision. Meanwhile, it is used as the decoder for upsampling. To strengthen the edge supervision of sea-land segmentation and predict sea-land masks and edges more accurately, an additional edge supervision branch is introduced into the decoder to predict the sea-land edges of the SAR image. The features of the two branches are interactively concatenated. This is because edge prediction and mask prediction are dual problems, and this feature interaction design can help improve the prediction accuracy of both. Considering that the edge feature is a shallow feature, it is necessary to make full use of the rich shallow information in the SAR image. Thus, this paper presents a relatively independent global CNN branch that directly handles the original image. It consists of two downsampling and two upsampling stages. The output of the last downsampling is added to the corresponding feature map of the encoder as the input of the additional boundary supervision branch. Meanwhile, the output of the last upsampling is added to the corresponding feature map in the decoder to predict the final mask. In addition, this paper presents a self-attention mechanism with a gating strategy into each basic transformer block. It helps the transformer architecture to better learn positional bias on small-scale SAR image datasets.

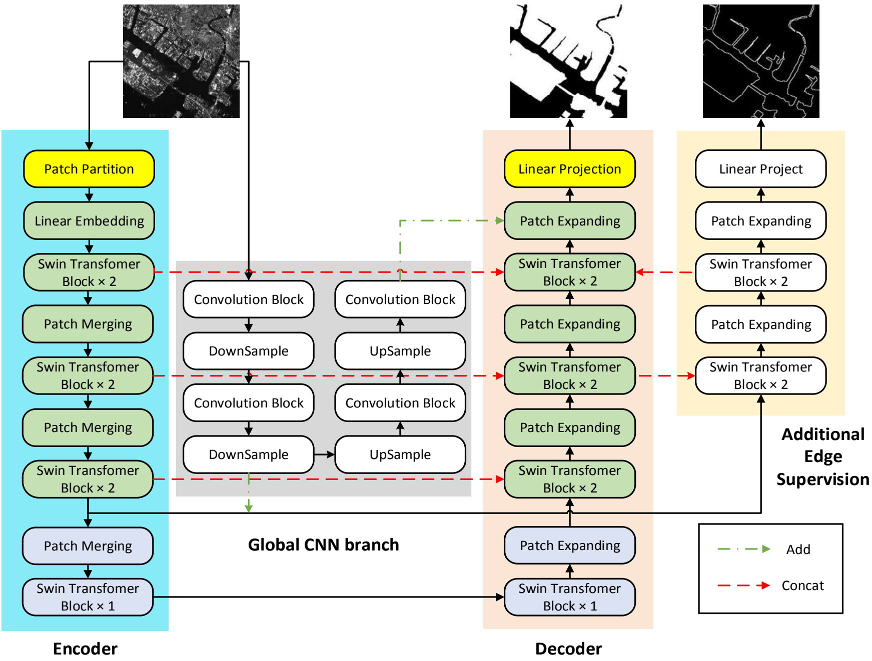


Fig. 1 The structure of the network model proposed in this paper

A. The Self-Attention Mechanism with a Gating Strategy

The basic unit of each down-sampling or up-sampling stage is composed of two consecutive transformer blocks, as shown in Fig. 2. Each transformer block consists of a Multi-head Self Attention (MSA) module and a Multi-Layer Perceptron (MLP) module. Before entering these two modules, a Layer Normalization (LN) layer is used. Then, the residual connection is used after each MSA module and MLP module. In the first transformer block, the MSA is the Widows-MSA (W-MSA) module, which evenly divides the feature map into small non-overlapping patches and calculates self-attention in each patch. Since the information between each patch cannot be communicated with each other, Shifted Windows-MSA (SW-MSA) is introduced into the next connected block. In SW-MSA, the feature map is re-divided so that the information in each patch can interact. The two window division methods of the MSA structure are illustrated in Fig. 3.

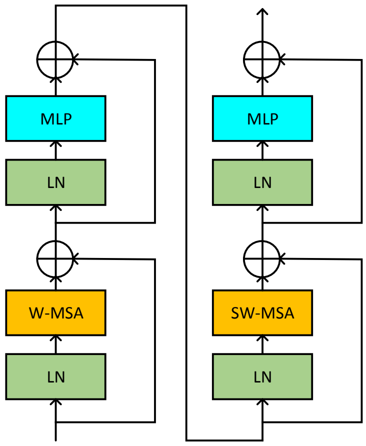


Fig. 2 Two consecutive transformer blocks

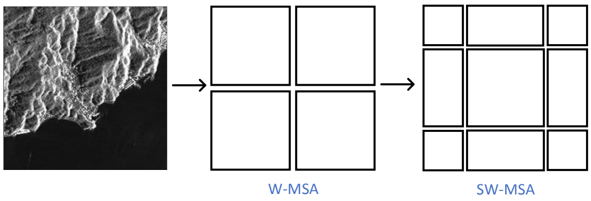


Fig. 3 Two window division methods of the MSA structure

In the MSA module, q(query), k(key), and v(value) are used to obtain the self-attention of the whole image, where q, k, and v are all learnable parameters. The output of the MSA module can be expressed as:

$$y_{ij} = \sum_{h=1}^{H} \sum_{w=1}^{W} soft\max(\frac{q_{ij}^{T} k_{hw}}{\sqrt{d}} + B) v_{hw} \qquad (1)$$

where d is the dimension of q and k, and B is the relative position bias.

Considering that SAR image datasets are relatively small, relative position bias is difficult to learn and long-distance features cannot effectively be captured, this paper proposes a self-attention mechanism based on a gating strategy. It can adaptively adjust the bias of query, key, and value through the gate control strategy to capture more accurate relative position bias information. The structure of the proposed self-attention mechanism is shown in Fig. 4, and the output can be expressed as:

$$y_{ij} = \sum_{w=1}^{W} softmax(\frac{q_{ij}^{T} k_{iw} + G_{Q} q_{ij}^{T} r_{iw}^{q} + G_{K} k_{iw}^{T} r_{iw}^{k}}{\sqrt{d}} + B)(v_{iw} + G_{V} r_{iw}^{v}) \qquad (2)$$

where r^q, r^k, and $r^v \in R^{W \times W}$ are the expansion of the axial attention mechanism to strengthen the ability to capture non-local information; G_Q, G_K, and $G_V \in R$ are learnable parameters. If the relative position bias can be accurately learned, the gating mechanism will assign a lower weight coefficient.

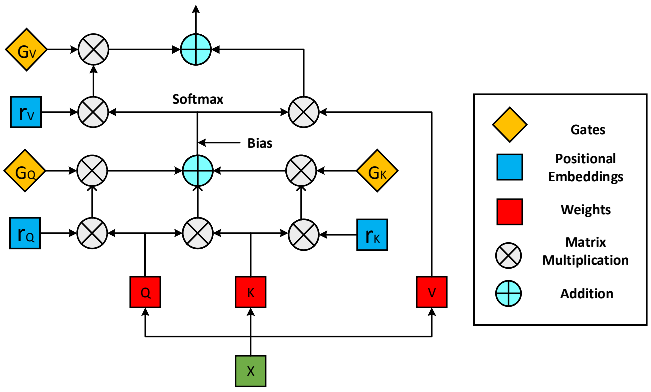


Fig. 4 The diagram of the proposed self-attention mechanism with a gating strategy

B. Additional Edge Supervision Branch

This paper presents an additional edge supervision branch to supervise the sea-land segmentation mask and predict the sea-land edge maps at the same time. It helps alleviate the problem of inaccurate sea-land edges caused by the scattering characteristics of SAR images. Edge features are shallow features, so it is necessary to make full use of the rich shallow feature information in the multi-scale features provided by the encoder.

Because the transformer architecture cannot learn the position bias on a small-scale SAR image dataset effectively, it is easy to reduce the segmentation accuracy. To address this issue, this paper presents a global CNN branch. As shown in Fig. 1, the global CNN branch has a similar structure to the encoder and decoder, but the difference is that this branch can ensure full utilization of the rich shallow features. So, only two downsampling stages are performed in the encoder, and two upsampling stages are performed in the decoder. The feature map of the last downsampling operation is added to the feature extracted from the backbone and sent to the additional edge supervision branch. This can not only alleviate the problem of accuracy loss caused by the position bias of the transformer in a small dataset but also make full use of the shallow feature information.

In the additional edge supervision branch, the fused feature maps are sent to two transformer blocks and then concatenated with the corresponding feature maps in the mask branch. Then, the concatenated result is taken as input and sent to the additional edge supervision branch. After the calculations of one up-sampling stage and two transformer blocks, the obtained features are concatenated with the features of the mask branch, and the concatenated result is used as input to the mask prediction branch. Considering that the prediction mask and the prediction edge are dual problems, this design of the two prediction branch feature interactions can supervise and learn from each other to achieve better segmentation performance. The design of the additional edge supervision is shown in Fig. 1, and the effectiveness of this design will be proven by the following experiments.

C. Loss Function

Since the method proposed in this paper needs to predict the sea-land mask and edge at the same time, the loss function consists of two parts: the loss of the sea-land mask and the loss of the sea-land edge. The overall loss function is presented as:

$$Loss_{total} = Loss_{mask} + Loss_{boundary} \tag{3}$$

For the sea-land mask loss, this paper directly uses the cross-entropy loss function as follows:

$$Loss_{mask}(p^m, y^m) = -\frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W} y_{hw}^m \cdot \ln(p_{hw}^m) + (1-y_{hw}^m)\cdot \ln(1-p_{hw}^m) \tag{4}$$

where p^m indicates the probability that a certain pixel is predicted to be land, and y^m indicates the true label of the pixel.

For the sea-land edge loss, this paper presents the dice loss based on the cross-entropy loss. This is because the edge feature information is difficult to learn. The cross-entropy loss alone cannot solve the problem of the imbalance of positive and negative samples in the edge label, and the dice loss can solve this problem. The loss function is presented as follows:

$$Loss_{CE}(p^b, y^b) = -\frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W} y_{hw}^b \cdot \ln(p_{hw}^b) + (1-y_{hw}^b)\cdot \ln(1-p_{hw}^b) \tag{5}$$

$$Loss_{Dice}(p^b, y^b) = 1 - \frac{2\sum_i^{H\times W} p_i^b y_i^b + \varepsilon}{\sum_i^{H\times W}(p_i^b)^2 + \sum_i^{H\times W}(y_i^b)^2 + \varepsilon} \tag{6}$$

$$Loss_{boundary} = Loss_{CE} + Loss_{dice} \tag{7}$$

where p^b indicates the probability that a certain pixel is predicted to be the edge of sea-land; y^b presented the true label of the pixel; ε indicates the smoothing term, and it is set to 1 to prevent the loss value from overflowing.

III. Experiment Results

In this study, the experiments are conducted on the Ubuntu 16.04 operating system; Pytorch is used as the deep learning framework, and the programs are written on the Pycharm software platform. The experimental hardware consists of an Inter(R) i5-10400F CPU, a GeForce GTX 2080 Ti GPU, and 16 GB memory. Besides, CUDA 11.1 is used to manage the GPU for training acceleration.

A. Dataset

The original data are obtained from the image data taken by the Gaofen-3 satellite at the port. The original data are manually labeled with sea-land mask labels. Then, the original image and the mask label are sliced into 256×256 small images, and the images that are all land or all sea area are removed. Finally, 1200 small images are filtered out, with differences in the proportion of sea and land. This prepared SAR image dataset includes simple and complex coastline scenes of the port and the scenes of many islands, which ensures the diversity of the dataset.

The prepared dataset is divided into a training set and a test set at the ratio of 7:3. Meanwhile, it is ensured that the data distribution of the training set and the test set are roughly the same. During the training process, the Adam optimizer is used for parameter replacement. In the experiment, the initial learning rate is set to 0.0001. All models are trained from scratch for 3000 epochs, and the training batch size is 4. During the training process, the model is saved every 200 epochs until the training is completed.

B. Evaluation Indexes

In this study, the mean intersection of union (MIoU), edge precision (EP), overall precision (OP), and F1 score are used to quantitatively evaluate the segmentation performance of the proposed method. The calculation formulas of the above evaluation indexes are as follows:

$$MIoU = \frac{A\cap B}{A\cup B} = \frac{TP}{TP+FP+FN} \tag{8}$$

$$EP = \frac{TP_{edge}}{TP_{edge}+FP_{edge}} \tag{9}$$

$$OP = \frac{TP}{TP + FP} \quad (10)$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (11)$$

where TP and FP respectively represent the predicted number of true-positive and false-positive samples; TP_{edge} and FP_{edge} respectively represent the true-positive and false-positive samples of the sea-land edge.

C. Comparative Experiment

The experiment can be divided into three parts, including the comparison of the performance between the transformer architecture and the CNN architecture, the verification of the effectiveness of the improved network structure, and the comparative analysis of the existing typical networks. The experimental results of each part are given below.

Performance Comparison of Different Network Architectures

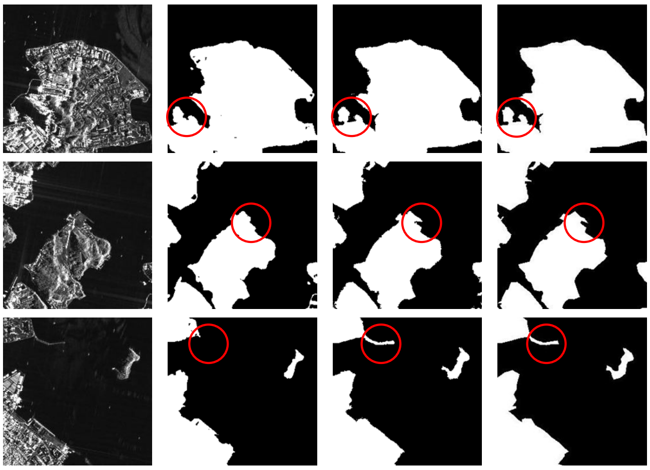
This paper uses a framework that has a similar structure to the U-Net for the entire network, in which the transformer architecture uses swin-transformer as the backbone, and the CNN architecture uses ResNet34 as the backbone. It can be seen from Table I that there is little difference in the MIoU and OP of the CNN architecture and the transformer architecture under the same conditions, but the transformer architecture improves the edge precision by 8.98%. Compared to the CNN architecture, the transformer architecture can better overcome the inherent bias to a certain extent and make a finer segmentation in the edge area of the sea and land.

TABLE I

PERFORMANCE COMPARISON OF DIFFERENT ARCHITECTURES

Backbone	Structure	MIoU	EP	OP	F1
ResNet34	U-Net	94.68	63.05	95.24	97.27
Swin-Transformer	U-Net	94.73	72.03	95.98	97.29

Fig. 5 shows some of the masks for the sea-land segmentation obtained by the two architectures. It can be seen that the sea-land mask of the transformer architecture has more accurate segmentation details than the CNN architecture, but it is not as good as the CNN architecture in terms of edge continuity and smoothness.



(a)Original image (b)CNN (c)Transformer (d)GroundTruth

Fig. 5. The comparison of the masks for the sea-land segmentation results of different architectures

The Effectiveness of Network Structure Improvement

The improved structure is compared with the baseline swin-unet (Transformer architecture). The improvement includes the attention mechanism with the gating strategy and the additional edge supervision branch.

a) The Effectiveness of the Self-Attention Mechanism with the Gating Strategy

Related research has proved that it is difficult for the transformer architecture to learn relative position bias on small-scale datasets. Therefore, this paper presents the attention mechanism with a gating strategy to adaptively control the weight coefficients of the learned information to improve segmentation accuracy. Taking Swin-UNet as the baseline, the effectiveness of introducing the self-attention mechanism with gating strategy is investigated. The results are shown in Table II. Compared with the baseline, the MIoU, EP, OP, and F1 of the model that only introduces the self-attention mechanism with the gating strategy are improved. Meanwhile, the improved method proposed in this paper can facilitate the learning of relative position bias, thereby better processing the information of sea-land edges. The edge precision of sea-land segmentation is increased by 1.26%.

TABLE II

THE EVALUATION OF THE NETWORK MODEL WITH DIFFERENT IMPROVEMENTS

Method	MIoU	EP	OP	F1
Swin-UNet (baseline)	94.73	72.03	95.98	97.29
Only with gated strategy	95.13	73.29	96.21	97.72
Edge supervision	95.62	74.28	96.01	97.61
Global CNN branch + edge supervision	96.17	80.99	96.74	97.94
Ours	96.36	84.54	99.74	98.05

b) The Effectiveness of the Additional Edge Supervision Branch

In Fig. 6, part of the data in the test set is taken to show the segmentation results and edge prediction results. Fig. 6 (a) shows the original image; Fig. (b) shows the result of sea-land segmentation without additional edge supervision; Fig. 6 (c) shows the edge supervision without global CNN branch, and Fig. 6 (d) shows the edge supervision fused with the global CNN branch.

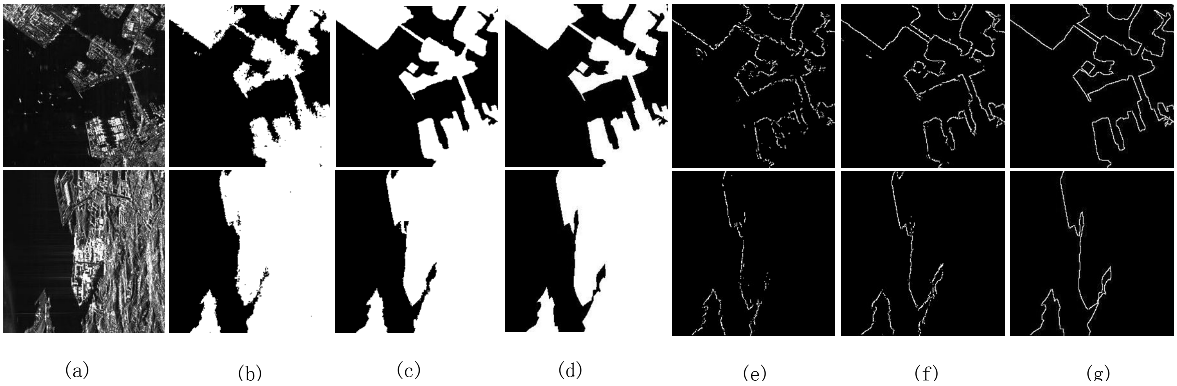


Fig. 6 The effectiveness of introducing the additional edge supervision

It can be seen from Fig. 6 that no matter whether the additional edge supervision is introduced, there are always some local burrs on the edges of the sea-land segmentation. The segmentation result is poor in continuity and not smooth enough. This may be an inherent bias of the transformer architecture, but it still has obvious advantages. Previous experimental results indicate that the transformer architecture can perform more refined processing on the sea-land edges than the CNN architecture and greatly improve the edge accuracy. The method proposed in this paper incorporates the global CNN branch into the additional edge supervision to ensure that the edge is smooth and close to the real label. Figs. 6 (e), (f), and (g) show the predicted edges and the corresponding true labels under the two additional edge supervisions.

It can be seen from Fig. 6 that after the global CNN branch is incorporated into the edge supervision, the edge of the segmentation is smoother and has fewer line breaks. This indicates the effectiveness of integrating the global CNN branch. Since the prediction edge and prediction mask are dual problems, it is reasonable that the introduction of additional edge supervision can improve the performance of sea-land segmentation. It can be seen in Table II that the use of the additional edge supervision can increase the MIoU by 0.89%, and the EP is significantly increased by 2.25%. Meanwhile, combined with the self-attention mechanism with the gating strategy, the final performance is further improved.

Comparative Analysis of the Existing Typical Networks

To more fully prove the segmentation performance of the method proposed in this paper, typical methods such as U-Net, Deeplabv3+, Swin-UNet, and MedT are selected for comparison. Taking three SAR images with different sea-land ratios as typical representatives, the segmentation results are shown in Fig. 7. Moreover, the results of the evaluation indexes of different typical networks are listed in Table III.

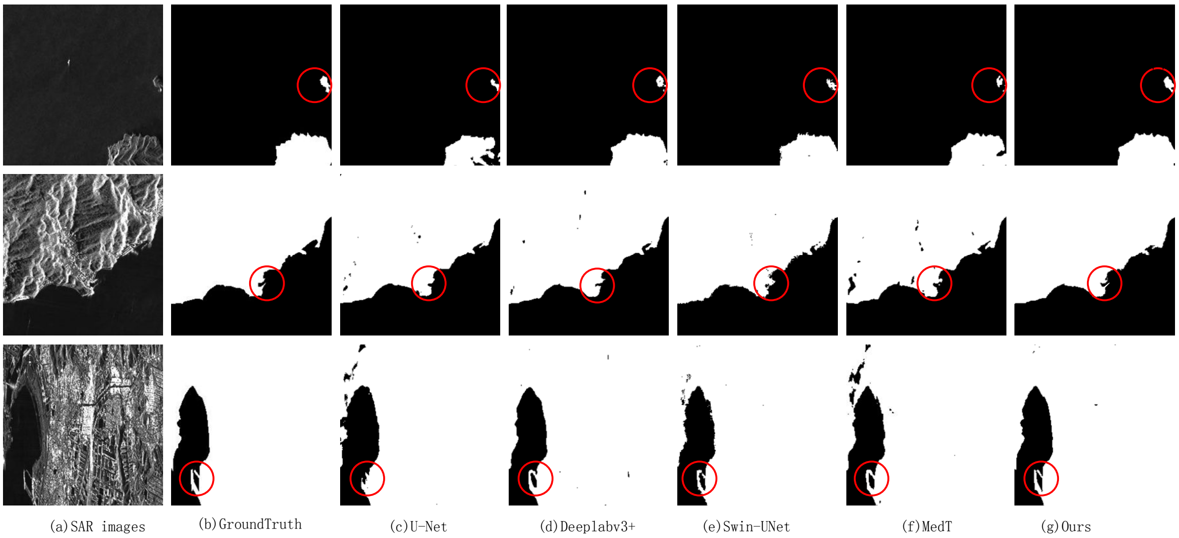


Fig. 7 Segmentation results of different typical networks

TABLE III

EVALUATION INDEXES OF DIFFERENT TYPICAL NETWORKS

Method	MIoU	EP	OP	F1
U-Net	94.68	63.05	95.24	97.27
Deeplabv3+	95.81	72.74	96.29	97.85
Swin-UNet	94.73	72.03	95.98	97.29
MedT	95.52	72.80	96.18	97.71
Ours	96.36	84.54	99.74	98.05

Among the methods taken for comparison, U-Net and Deeplabv3+ are based on the CNN architecture, while Swin-UNet and MedT are based on the transformer architecture. It can be seen from the segmentation results in Fig. 7 that the sea-land edge segmentation results of U-Net and Deeplabv3plus are relatively smooth and have fewer discrete points, but the sea-land edges are not segmented well for a relatively narrow land area. Swin-UNet and MedT can better segment the sea-land edges, but the number of voids in the land is not reduced. Besides, the presence of burrs on the sea-land edges leads to relatively many isolated and bright pixels nearby. The method proposed in this paper is superior to other algorithms. It achieves better connectivity in the segmentation results, greatly reduces the central voids in the land, processes the details of the sea and land edges more accurately, eliminates locally isolated bright pixels on the sea and land edges, and makes the sea-land edges smooth. The superiority of this algorithm is mainly attributed to the additional edge supervision designed in this paper. Moreover, it can also be seen from Table III that the proposed method in this paper improves the EP and MIoU significantly. The EP and MIoU respectively reach 84.54% 96.36%, which are better than those of other algorithms.

IV. CONCLUSION

This paper presents a sea-land segmentation method based on the transformer architecture to strengthen edge supervision. This method introduces a self-attention mechanism with a gating strategy to solve the problem that transformer is difficult to learn relative position bias in small-scale data. Meanwhile, it introduces additional edge supervision to greatly improve the accuracy of edges during sea-land segmentation. Based on the Gaofen-3 satellite image data, the performance of this method is compared with that of mainstream segmentation algorithms. The experimental results show that using the method proposed in this paper, the MIoU, EP, OP, and F1 score can respectively reach 96.36%, 84.54%, 99.74%, and 98.05%. The research results have high practical values and can be applied to other related tasks such as ship inspection.

REFERENCES

[1] Liu Sitong, Cheng Hong, Sun Wenbang, and Yu Guang, “Studies of sealand segment methods oriented to targets on the sea,” Electronic Design Engineering, vol. 22, no. 15, pp. 96-100, August 2014.

[2] J. S. Lee and I. Jurkevich, “Coastline detection and tracing in SAR images,” IEEE Trans. Geosci. Remote Sens. 28(4), 662–668 (1990).

[3] D. C. Mason and I. J. Davenport, “Accurate and efficient determination of the shoreline in ERS-1 SAR images,” IEEE Trans. Geosci. Remote Sens. 34(5), 1243–1253 (1996).

[4] D. Cheng et al., “Efficient sealand segmentation using seeds learning and edge directed graph cut,” Neurocomputing 207, 36–47 (2016).

[5] Li Jianwei, Qu Changwen, and Shao Jiaqi. Ship detection in SAR images based on an improved faster R-CNN (C). 2017 SAR in Big Data Era: Models, Methods and Applications, Beijing, China, 2017: 1–6. doi: 10.1109/BIGSA RDATA.2017.8124934.

[6] G. Liu, Y. Zhang, X. Zheng, X. Sun, K. Fu, and H. Wang, “A new method on inshore ship detection in high-resolution satellite images using shape and context information,” IEEE Geosci. Remote Sens. Lett., vol. 11, no. 3, pp. 617–621, Mar. 2014.

[7] He You, Huang Yong, Guan Jian, and Chen Xiaolong, “An overview on radar target detection in sea clutter,” Modern Radar, vol. 36, no. 12, pp.

1-9, December 2014.

[8] Ma Xiaoli, “Fast segmentation method of clutter scene of land and sea,” Master degree, Xidian University, Xi’an, China, December 2014.

[9] Cheng D, Meng G, Cheng G, et al. SeNet: Structured Edge Network for Sea-Land Segmentation. IEEE Geoscience and Remote Sensing Letters, Papers 14(2), 247-251(2017).

[10] Cheng D, Meng G, Member, et al. FusionNet: Edge Aware Deep Convolutional Networks for Semantic Segmentation of Remote Sensing Harbor Images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, Papers (99),1-15 (2017).

[11] R. Hansch, O. Hellwich, and X. Wang, “Graph-cut segmentation of polarimetric SAR images,” in IEEE Int. Geoscience Remote Sensing Symp. (IGARSS), 1733–1736 (2014).

[12] Houxy, Xuf. Hybrid strategy for precise sea-land segmentation GF-3SAR images (J). Chinese journal of radio science, 2019, 34(6): 798-805. (In Chinese). DOI: 10.13443/j. cjors. 2019042801.

[13] Ma Xiaoli, “Fast segmentation method of clutter scene of land and sea,” Master degree, Xidian University, Xi’an, China, December 2014.

[14] X. Zhao, Y. Jiang, and W. Wang SAR image segmentation based on analyzing similarity with clutter spatial patterns Electron. Lett., vol. 52, no. 21, pp. 1807–1809, 2016.

[15] Ruirui Li, Wenjie Liu, Lei Yang, Shihao Sun, Wei Hu, Fan Zhang, and Wei Li, “Deepunet: A deep fully convolutional network for pixel-level sea-land segmentation,” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2018.

[16] Zhihong Pan, Hao Dou, Jiaxing Mao, Min Dai, and Jinwen Tian. MIFNet: Multi-Information Fusion Network for Sea-Land Segmentation. In ICAIP (2018).

[17] H. Lin, Z. Shi, and Z. Zou, “Maritime semantic labeling of optical remote sensing images with multi-scale fully convolutional network,” Remote sensing, vol. 9, no. 5, p. 480, 2017.

[18] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in Medical Image Computing and Computer-Assisted Intervention, Cham, 2015, pp. 234–241.

[19] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla, “SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017.

[20] L. C. Chen, G Papandreou, I Kokkinos, K Murphy, and A. L. Yuille, “DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs,” IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 40, no. 4, pp. 834–848, 2018.

[21] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel, “Understanding the Effective Receptive Field in Deep Convolutional Neural Networks”, arXiv preprint arXiv:1701.04128.

[22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” arXiv preprint arXiv:2010.11929.

[23] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, et al, “End-to-End Object Detection with Transformers,” arXiv preprint arXiv:2005.12872v3.

[24] Ze Liu, Yutong Lin, Yue Cao, Han Hu, et al, “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows,” in CVPR (2021).

LFM signal perception based on Wavelet transform and Time-Frequency Technology

Xingcai Wang

***Abstract*—**Linear frequency modulation signals are a common modulation method for low intercept probability radar signals, a spread-spectrum modulation technique that does not require pseudo-random coding sequences and has been widely used in radar and sonar technology due to its large time-frequency product. In order to improve the perception of LFM signals in a low SNR environment, this study proposes a time-frequency analysis method for LFM signals based on segmentation denoising, wavelet transform denoising, and Choi-Williams Distribution. The results show that the method has good performance and feasibility under low SNR conditions and can exhibit clear time-frequency characteristics of the LFM signal at a SNR of -21dB. Finally, combined with deep learning, GoogLeNet is used as the training network and the time-frequency image as the training sample, which achieves a good signal detection probability. The detection probability is greater than 90% when the SNR is greater than -18dB, and the overall detection probability is better than other detection network models.

***Keywords*—**Linear frequency modulation signal, Choi-Williams distribution, Segmentation denoising, Wavelet transform denoising, time-frequency analysis, Deep learning.

I. Introduction

In recent years, with the increasing complexity of the battlefield electromagnetic environment and the wide application of new radar systems, it has become more and more difficult to detect and analyze signals in a low signal-to-noise ratio (SNR) environment. LPI radars avoid the interception and detection of their signals by non-cooperative electronic reconnaissance aircraft by emitting waveforms that are modulated by special modulation [1,2]. LFM as a common modulated signal type in LPI radars, its frequency time variation and large time-frequency product characteristics make it widely used in radar and sonar technology, so the research on the perception of LFM signals is of great significance. In order to improve the perception ability of electronic reconnaissance, LPI radar signal perception under low SNR has gradually become a research focus of the electronic warfare system. In the LPI radar detection and classification problem, the traditional FFT-based signal detection method is no longer applicable because it cannot capture the details of the signal changes. In recent years, based on the time-frequency image, the time-frequency analysis technique has been widely used in low probability of intercept radar signal detection and has achieved good results. The time-frequency analysis method can describe the energy intensity and distribution characteristics of signals simultaneously, which is an effective and direct method to analyze non-stationary signals. Time-frequency analysis (TFA) techniques such as windowed short-time Fourier Transform (STFT), fractional Fourier transform, Wigner-Ville distribution (WVD) and Choi-Williams distribution (CWD) are commonly used to describe the distribution characteristics of signals in the time and frequency domain [3].

Over the years, researchers have done a lot of work on LPI radar waveform detection and modulation classification. In [4], various LPI modulated waveforms embedded in noise with a SNR of -15 dB are classified and featured by using Wigner-Ville distribution and fractional Fourier transform. In [5], the author realizes the detection and classification of LPI waveform modulation with a SNR of -20dB by selecting the short-time Fourier transform of appropriate window length and type as the time-frequency analysis method. In [6], the author uses fractional Fourier transform to construct the classifier, and achieves an overall recognition rate of 94.17% under the condition of 0dB.

With the development of wavelet transforms, widely used in the field of signal analysis, as [7]. The wavelet transform has the advantages of low entropy, sparse distribution of wavelet coefficients, and multi-resolution, which can focus on the details of the signal for time-frequency domain analysis, making it show good denoising ability in signal processing. With the development of time-frequency analysis technology, Choi-Williams distribution (CWD) has been widely used in a variety of engineering fields, including image analysis, target detection, and analysis of non-stationary signals. In non-stationary signals, CWD adopts an exponential kernel function to suppress and eliminate the effects of cross terms, resulting in a time-frequency transform image with good T-F resolution compared with WVD.

In this study, a CWD time-frequency detection method for LFM signals based on wavelet transform denoising (CWDW) is proposed. Finally, the LFM signal is detected by combining wavelet transform denoising and CWD, and finally the identification of the LFM signal under the condition of an SNR of -21dB is realized.

This paper is organized as follows: system signal models, wavelet transform denoising, and time-frequency analysis techniques are described in Section II; signal parameters, denoising process, time-frequency image simulation results, and detection probability are described in Section III; and finally, conclusion is given in Section IV.

Xingcai Wang is with Research Institute of Electronic Science and Technology, University of Electronic Science and Technology of China, Chengdu, China (e-mail: 202021230117@std.uestc.edu.cn).

II. SIGNAL MODEL AND TFA TECHNIQUES

In this section, the LFM signal waveform, wavelet transform denoising, and time-frequency analysis techniques used are described.

A. LFM Signal Model

As a typical non-stationary signal, linear frequency modulation (LFM) signal is widely used in radar, sonar, communication, and other information systems [8], such as Doppler echo signals and moving target detection of synthetic aperture radar, which is essentially the detection of linear frequency modulation signals, and can also be expressed by the following formula:

$$x(t) = Arect\left(\frac{t}{T}\right)\exp\left(j2\pi\left(f_c t + \frac{kt^2}{2}\right)\right) \tag{1}$$

where $k = B/T$ indicates the represents frequency modulation slope, T is the pulse period of the LFM signal, B is the modulation bandwidth of the signal, f_c is the signal carrier frequency, and$rect(t/T)$is a rectangular signal.

In engineering applications, the intercepted LPI signal is affected by the environment and system additive noise, and the actual received signal is as follows:

$$y(t) = x(t) + n(t) \tag{2}$$

Where n(t) symbolizes the complex additive white Gaussian noise (AWGN).

B. Choi-Williams Distribution

T-F analysis techniques are an important method for processing non-stationary signals. The bilinear generalized exponential kernel function $\emptyset(\theta,\tau) = e^{-\theta^2\tau^2/\sigma}$ is used in the Choi-Williams distribution (CWD) to suppress and eliminate the effects of the cross terms, so that the time-frequency transform image has good T-F resolution, where θ and τ represent frequency domain lag and time lag, respectively. CWD functions can be expressed as:

$$CWD_y(t,\omega) = \int_{\tau=-\infty}^{+\infty} e^{-j\omega\tau} \int_{\mu=-\infty}^{+\infty} \sqrt{\frac{\sigma}{4\pi\tau^2}}\, e^{-\frac{(\mu-t)^2}{4\tau^2}} \cdot y\left(\mu + \frac{\tau}{2}\right) y^*\left(\mu - \frac{\tau}{2}\right) d\mu d\tau \tag{3}$$

where t is the time variable, ω is the angular frequency variable ($\omega = 2\pi f$), and σ is the positive scaling factor.

C. Wavelet Transform Denoising

Noisy one-dimensional signal models can be expressed as:

$$y(k) = x(k) + n(k), k = 0, 1, \cdots, n-1 \tag{4}$$

Where $y(k)$is a noised-signal, $x(k)$is a useful signal, $n(k)$ is a noise signal, in engineering applications $x(k)$ is usually a signal with a certain frequency range, noise is usually expressed as a high-frequency signal, In this paper, the following steps are used to de-noise the signal.

Step 1: Wavelet decomposition of one-dimensional signals. Select a wavelet and determine the level of decomposition, and then perform the decomposition calculation.

Step 2: Threshold quantization of wavelet decomposition high frequency coefficients. Select a threshold for the high-frequency coefficients at each decomposition scale for soft threshold quantization.

Step 3: One-dimensional wavelet reconstruction. One-dimensional wavelet reconstruction is performed according to the lowest layer low frequency coefficient and each high frequency coefficient of the decomposition.

III. EXPERIMENTS AND DISCUSSION

Based on the above time-frequency analysis technology, the LTM signals under different SNRs are simulated in this section, and the time-frequency images under different SNRs are obtained, which are used to determine whether the signal exists or not.

A. Parameters Setting

As a typical non-stationary signal, the linear frequency modulation (LFM) signal is widely used in radar, sonar, communication system and other information systems [8], it is universal in practical engineering applications. The research of this paper aims at the perception of LFM signal at low SNR, and the signal parameters are shown in the following TABLEI:

TABLE I
WAVEFORM PARAMETERS

Waveform	BW	PW	f_c	f_s
LFM	10KHz	0.01s	0Hz	1MHz

B. Denoising Processing

Signals in the real environment often have strong noise interference, the SNR is below 0, in order to improve the effect of signal denoising, this paper takes two steps to denoising.

Step 1: In this scheme, the noisy signal is first filtered in segments, for most non-stationary signals, in a short time interval $(t_0, t_0 + T_0)$, each modulation signal can be approximated as a sinusoidal signal, so use this feature to denoise the signal for the first time.

After segmentation of the signal, the m-segment signal can be expressed as:

$$\begin{gathered} y_i(t) = A\exp[j(2\pi f_0 t + \theta_0)] + n(t) \\ iT_0 \le t \le (i+1)T_0 \end{gathered} \tag{5}$$

Discrete sampling can be expressed as:

$$\begin{gathered} y_i(m) = \quad A\exp[j(2\pi f_0 m\Delta t + \varphi)] + n(m) \\ , \quad i(M_0 - 1) \le m \le (i+1)(M_0 - 1) \end{gathered} \tag{6}$$

where the $M_0$is the number of sample points of the signal segment, the sampling interval is $\Delta t = T/M$, and M is the number of sample points of the signal.

The specific steps are as follows[9]:

(1) Do M_0 point DFT on $y_i(m)$ and get $R_i^{'}(k) = DFT[y_i(m)]$.

(2) Design a band-pass filter with transmission characteristics as follows:

$$H(k)=\begin{cases}1, & k_0-d\leq k\leq k_0+d\\ 0, & \text{other}\end{cases} \quad (7)$$

where k_0 is the $|X(k)|$ maximum line position, with d being the number of filter points.

(3) Order $R_i'(k)=H(k)R_i(k)$, then do M-spot IDFT on $R_i'(k)$, get $y_i'(n)=IDFT(R_i'(k))$.

(4) Combine each segmented reconstructed time domain signal into a new observation signal $y'(k)$.

Step2: For the noised-signal, db5 wavelet are used to decompose the signal at 5 layers(Fig.1), take the low frequency detail coefficient of layer 5, take the high frequency detail coefficient of each layer, choose an appropriate soft-threshold, and finally reconstruct the signal according to the low frequency coefficient of layer 5 and the high frequency coefficient of each layer to obtain a denoising signal. Fig.2 and Fig.3 show the low frequency details and high frequency details of the fourth layer, respectively.

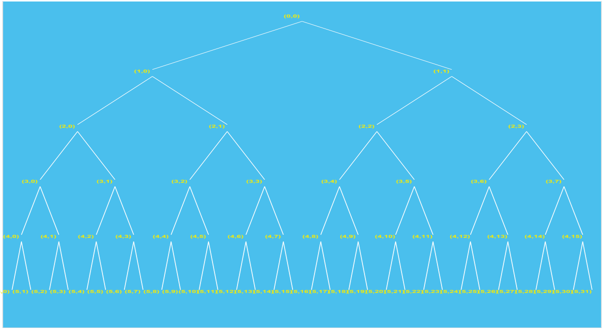


Fig.1 Tree decomposition

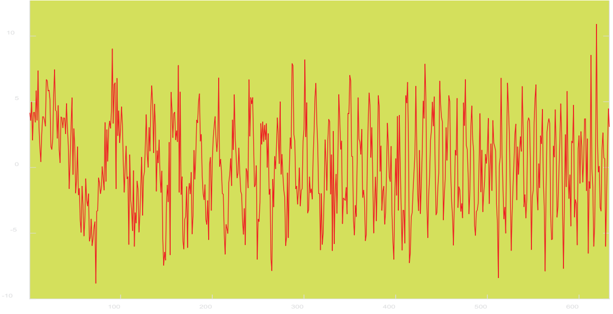


Fig.2 Low frequency component

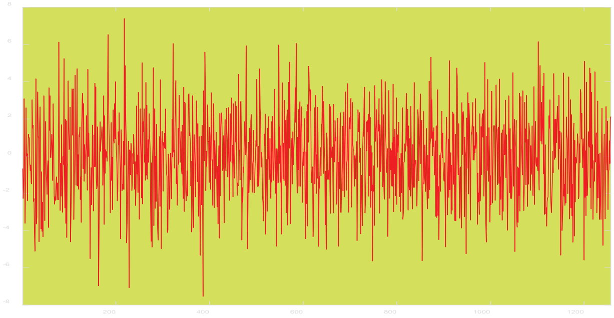


Fig.3 High frequency component

According to the above denoising steps, in order to achieve good denoising effect, the signals of 10000 data points are processed in segments, and the amount of data in each group is M_0 = 20, filter length d = 7, the signal is reconstructed after segmented filtering, and then the final denoised signal is obtained through wavelet denoising.

TABLE II

SNR COMPARISON

ORIGINAL SNR	SNR AFTER DENOISING
-10dB	-1.9dB
-15dB	-8.9dB
-20dB	-16dB

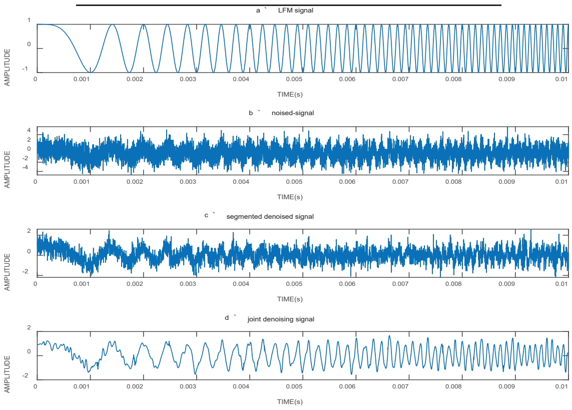
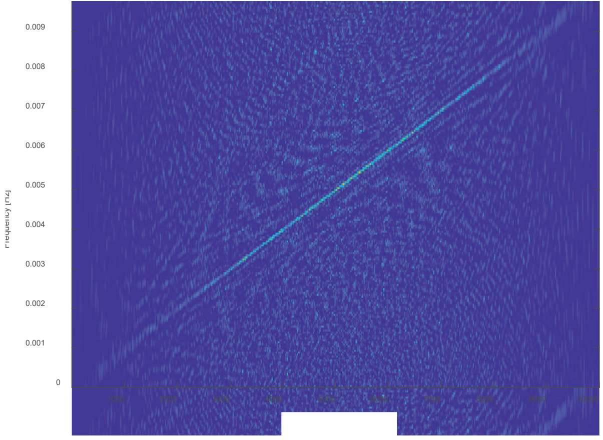


Fig.4. Signal comparison of different steps

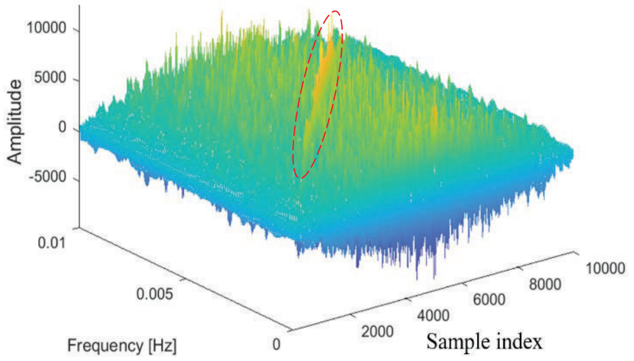
It can be seen from Fig.4 that after joint denoising, even when the SNR is negative, the signal hidden in the noise is extracted, which can meet the needs of a certain scene, and then the characteristics of the signal are found through time-frequency technology. TABLEII shows the SNR before and after denoising

C. *Time-frequency Distribution Image*

Under the above parameters, the LFM signal with a SNR of -18~-21dB was selected for simulation analysis. First, the signal is de-denoised by wavelet transformation, then the appropriate threshold is selected to achieve the best denoising effect, and then the time-frequency analysis is performed on the signal after denoising. In the CWD time-frequency simulation, in order to obtain the best time-frequency image, information entropy is adopted as an auxiliary method to continuously adjust the window type and length of time and frequency to obtain the minimum information entropy and finally obtain the best time-frequency diagram, as shown in the figure below:

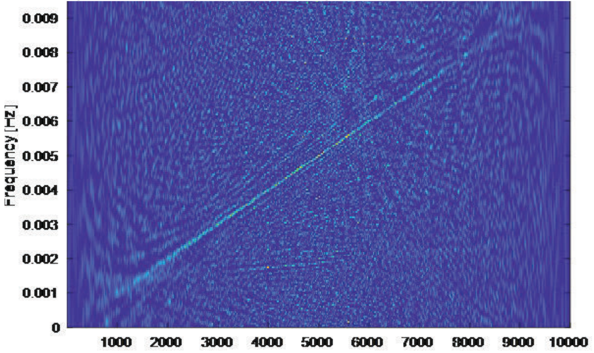


a. two-dimensional T-F image of CWDW

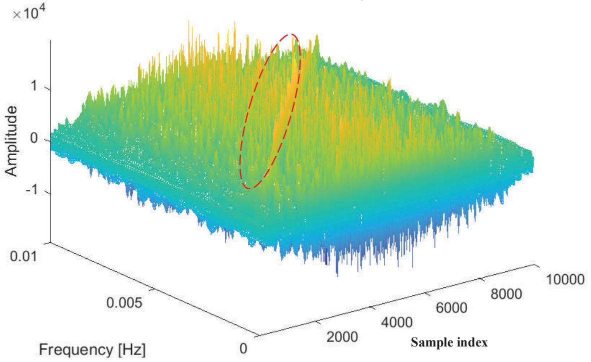


b. three-dimensional T-F image of CWDW

Fig.5 T-F images of LFM signals with SNR of -18dB

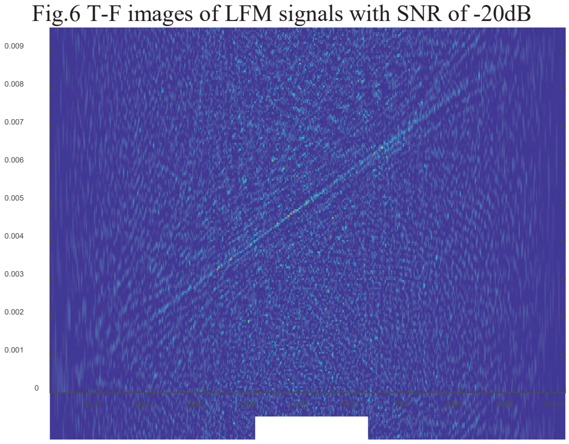


a. two-dimensional T-F image of CWDW

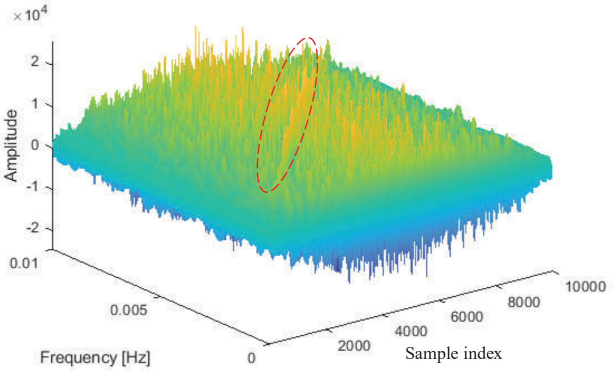


b. three-dimensional T-F image of CWDW

Fig.6 T-F images of LFM signals with SNR of -20dB



a. two-dimensional T-F image of CWDW



b. three-dimensional T-F image of CWDW

Fig.7 T-F images of LFM signals with SNR of -21dB

The simulation results show that CWD based on wavelet transform denoising has good detection ability for LFM signals. The low entropy and multi-resolution of wavelet denoising and the time-frequency aggregation of CWD make the time-frequency energy of the signal more concentrated. It can be clearly seen from Figs.5,6, and 7 that when the SNR is -18dB, -20dB, and -21dB, the frequency of the LFM signal changes linearly with time, which is consistent with the theoretical time-frequency characteristic distribution of the LFM signal. The simulation results show that the detection method adopted in this paper is still effective when the SNR of LFM signals is -18 ~ -21dB and the time-frequency energy is concentrated. The time-frequency distribution characteristics of LFM signals can be clearly seen in both two-dimensional and three-dimensional graphs. When the SNR is less than -21dB, the time-frequency image becomes more and more fuzzy, but the LFM signal characteristics can still be seen. After -22dB, signal detection becomes less noticeable. If combined with deep learning algorithms, better detection results will be achieved, which is also an important research direction for deep learning in signal detection.

D. Detection probability based on Googlenet network

With the development of deep learning networks, it brings new opportunities to signal processing. Using a deep learning networks, radar waveform can be recognized robustly from the time-frequency representation of WVD and CWD. For example, Wang et al. [10] designed a CNN, there are four conv layers in the cascade, and there are no series and skip connections, resulting in low learning efficiency. Kong et al. [11] Specified two conv layers and three fc layers in the architecture to significantly increase the network scale, which brings burden to efficient learning operation.Through the ingeniously designed module, GoogLeNet can make more efficient use of computing resources and extract more features with the same amount of computing, so as to improve the training results.

This section will combine the deep learning network to detect the LFM signal. GoogLeNet is used as the training network, and the time-frequency image data obtained by the method in Section C above is used as the training sample to construct the detection network for the LFM signal. Then, the time-frequency map data under different SNR are re obtained through experiments as test samples to obtain the detection probability under different SNR, as shown in Fig. 8. In Fig. 8, the recognition probability of the detection network in this paper is compared with that of Kong et al. [9] and Huynh-The et al. [3] (the minimum SNR of Kong and Huynh-The is - 20dB in their article and - 24dB in this study.). It can be seen from the figure that the detection network proposed in this paper achieves high recognition accuracy under low SNR and is better than that of the other two people. If the SNR is greater than - 18db, the detection probability is about 90% or more. It is worth noting that when the SNR is greater than - 10dB, the detection probability is basically close to 100%. However, as the signal-to-noise ratio decreases, the signal characteristics are drowned out by the noise, resulting in a sharp decrease in the probability of detection, to 76.47% at SNR=-20dB,56.86% at SNR=-21dB, 23.52% at SNR=-23dB, and 9.804% at SNR=-24dB.

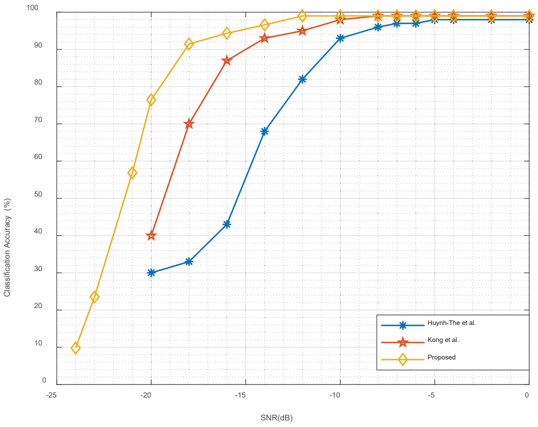


Fig.8 Recognition probability comparison

As can be seen from Fig. 8, although in the previous section, the time-frequency diagram can still show the basic characteristics of the LFM signal when the SNR is -20dB and -21dB. However, under such a low SNR, the recognition probability of the detection network is still seriously affected. But there is still a considerable detection probability.

IV. Conclusion

This study presents a time-frequency detection method of LFM signals based on CWD with segmentation denoising and wavelet transform denoising. This method combines the advantages of wavelet transform denoising and Choi-Williams distribution and uses information entropy as an auxiliary to select the appropriate parameters of time and frequency window, which makes the experiment get good detection results. Compared with other methods, this method has the following advantages: 1) It has good detection performance for LFM signals; 2) The algorithm has low complexity and fast detection speed. Finally, the experimental simulation results show that the method has good LFM signal detection ability under low SNR conditions and finally realizes the feature extraction of LFM signals with a SNR of -21dB. Combined with deep learning, GoogLeNet is used as the training network, and time-frequency image data is used as the training sample to obtain the LFM signal detection network, which achieves a good signal detection probability. When the SNR is greater than -18dB, the detection probability is greater than 90%, and the SNR is The detection probability is 76.47% at -20dB, and the detection probability is 56.86% when the SNR is -21dB.

References

[1] M. Gupta, G. Hareesh, A.K. Mahla, Electronic warfare: issues and challenges for emitter classification. Def. Sci. J. 61(3), 228–234 (2011). https:// doi. org/ 10. 14429/ dsj. 61. 529

[2] Aljaafreh and L. Dong, “An evaluation of feature extraction methods for vehicle classification based on acoustic signals,” in Networking,Sensing and Control (ICNSC), 2010 International Conference on, April 2010, pp. 570–575.

[3] T. Huynh-The, V. -S. Doan, C. -H. Hua, Q. -V. Pham, T. -V. Nguyen and D. -S. Kim, "Accurate LPI Radar Waveform Recognition With CWD-TFA for Deep Convolutional Network," in IEEE Wireless Communications Letters, vol. 10, no. 8, pp. 1638-1642, Aug. 2021, doi: 10.1109/LWC.2021.3075880.

[4] T. Ravi Kishore and K. D. Rao, "Automatic Intrapulse Modulation Classification of Advanced LPI Radar Waveforms," in IEEE Transactions on Aerospace and Electronic Systems, vol. 53, no. 2, pp. 901-914, April 2017, doi: 10.1109/TAES.2017.2667142.

[5] A. Gupta and A. A. Bazil Rai, "Feature Extraction of Intra-Pulse Modulated LPI Waveforms Using STFT," 2019 4th International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT), 2019, pp. 742-746, doi: 10.1109/RTEICT46194.2019.9016799.

[6] X. Cunxiang, Z. Limin and Z. Zhaogen, "Quasi-LFM radar waveform recognition based on fractional Fourier transform and time-frequency analysis," in Journal of Systems Engineering and Electronics, vol. 32, no. 5, pp. 1130-1142, Oct. 2021, doi: 10.23919/JSEE.2021.000097.

[7] Jie Xie, M. Towsey, P . Eichinski, Jinglan Zhang, and P . Roe, “Acoustic feature extraction using perceptual wavelet packet decomposition for frog call classification,” in e-Science (e-Science), 2015 IEEE 11th International Conference on, Aug 2015, pp. 237–242.

[8] Guo Y, Yang L. Method for parameter estimation of LFM signal and its application. IET Signal Proc 2019;13(5):538–43.

[9] HU Guo-bing, LIU Yu.Specific radar emitter recognition based on maximum-likelihood criterion. Systems Engineering and Electronics,2009,31(02):270-273.

[10] C. Wang, J. Wang, and X. Zhang, “Automatic radar waveform recog-nition based on time-frequency analysis and convolutional neural network,” in Proc. IEEE Int. Conf. Acoust. Speech Signal Process.(ICASSP), New Orleans, LA, USA, 2017, pp. 2437–2441.

[11] S. Kong, M. Kim, L. M. Hoang, and E. Kim, “Automatic LPI radar waveform recognition using CNN,” IEEE Access, vol. 6, pp. 4207–4219,2018.

Chaotic Sequence Noise Reduction and Chaotic Recognition Rate Improvement Based on Improved Local Geometric Projection

Rubin Dan, Xingcai Wang, Ziyang Chen

***Abstract*—**In this paper, a chaotic time series noise reduction improvement method based on the fusion of the local projection method, the wavelet transform, and particle swarm optimization is proposed to solve the problem of false recognition caused by noise in the recognition process of chaotic time series that contain noise. This method first uses phase space reconstruction to recover the original dynamical system characteristics and removes the noise subspace by selecting the nearby radius; then uses wavelet transform to remove D1-D3 high-frequency components to achieve maximum the retention of signal information, while simultaneously performing least-squares optimization via particle swarm optimization. The Lorenz system containing 30% Gaussian white noise is simulated and validated, and the phase space, signal-to-noise ratio, root mean square error, and K value of the 0-1 test method before and after noise reduction by the Schreiber method, the local projection method, the wavelet transform method, and the improved method are compared and analysed. This demonstrates that the improved method has a superior noise reduction effect than the other three methods. This further demonstrates the superiority of the enhanced approach. Last but not least, it is used to the Chengdu rainfall chaotic sequence for study, and the findings demonstrate that the revised approach can successfully decrease noise and increase the chaos detection rate.

***Keywords*—**Schreiber noise reduction, Wavelet transform, particle swarm optimization, 0-1 test method, chaotic sequence denoising.

I. Introduction

CHAOTIC sequences are sequences that study the evolutionary laws of systems based on time, and studying chaotic sequences reveals and restores the evolutionary laws of dynamical systems. Many sequences, such as sunspots, rainfall, and the financial stock market, exhibit a chaotic nature [1]-[3]. However, in a complex environment, the actual collected chaotic signals are filled with a lot of noise, and if chaos identification is performed in the original sequence, it not only takes a lot of time but also affects the accuracy and reliability of the algorithm. At the moment, the most important thing to figure out is how to get rid of the noise in a signal that is full of noise and chaotic, how to put the chaotic signal back together, and how to fix the dynamical system [4]-[6].

Many scholars at home and abroad have done research related to the noise reduction of noisy chaotic sequences. Grassberger et al. proposed a local projection method [7] with Schreiber's simple noise reduction method based on the phase space structure characteristics to maintain the integrity of chaotic dynamical systems for noise reduction [8]. Leontitsis et al. proposed an adaptive local projection noise reduction method from a feature point approach where the signal subspace is the useful feature points and the noise subspace is the useless feature points for noise reduction [9]. Mera et al. proposed a maximum likelihood criterion based on reducing the distance from points in the time series to the chaotic attractor [10], as Hussain et al. proposed a nonlinear adaptive denoising algorithm [11], which is based on dividing the time series into segments and taking the overlapping points of adjacent segments for K-order polynomial fitting to approximate the original series for noise reduction. David et al. used smooth orthogonal decomposition combined with an improved local projection method [12], and Karrech et al. proposed a multi-scale higher-order refinement of the improved local projection method to achieve effective noise reduction of chaotic sequences [13]. Zhenglong et al. proposed a neural network algorithm trained by using a particle swarm optimization algorithm for chaotic system identification [14]. Wu and Ma used a threshold selection method in thresholding the detailed part after boosting wavelet transform and combined it with particle swarm optimization to globally search for the optimal threshold [15].

II. Basic Theory

A. Local Projection Method and Wavelet Noise Reduction Principle

Let x_i be a noisy chaotic time series of the reaction chaotic system, where the clean sequence is y_i and the Gaussian white noise is n_i, which can be expressed as$x_i = y_i + n_i$,i ∈ [1, n] . The embedding dimension m and delay time τ of this chaotic time series are determined by the c-c method, and the phase space is reconstructed according to Takens' principle, which can be expressed as

$$X_i = (x_i, x_{i+\tau}, \ldots, x_{i+(d-1)\tau}) \tag{1}$$

Rubin Dan, Xingcai Wang, and Ziyang Chen are with Research Institute of Electronic Science and Technology, University of Electronic Science and Technology of China, Chengdu, China (e-mail: 202022230119@std.uestc.edu.cn, 202021230117@std.uestc.edu.cn, bandityoung@foxmail.com).

$$X_i = (x_i, x_{i+\tau}, \ldots, x_{i+(d-1)\tau}), \quad 1 < i < n - (d-1)\tau$$

We select the neighborhood radius ∈, for each $x_i$find a set $U_i^{\in}$ let all points x_j in the neighborhood satisfy: $\|x_j - x_i\| < \in$,at this time replace x_i with the average value of $x_j$satisfying in $U_i^{\in}$:

$$x_i^{corr} = \frac{1}{|U_i^{\epsilon}|} \sum_{U_i^{\epsilon}} x_j \tag{2}$$

$$W_i = X_i - X_i^{corr} \tag{3}$$

Equation (3) generates the neighborhood matrix W_i and then calculates the covariance matrix $C_i = (W_i)^T(W_i)$, which is formed of W_i. R represents C_i diagonal weight matrix. The eigenvalues and eigenvectors of C_i are determined, and the projection matrix Q_i is formed by the eigenvector e_j with p minimal eigenvalues.

$$Q_i = \sum_{j=d-p+1}^{d} e_j e_j^T \tag{4}$$

The projection of W_i into Q_i orthogonal projection space, followed by the correction of X_i.

$$\widehat{X}_i = X_i - R^{-1} \sum_{j=d-p+1}^{d} e_j \left(e_j^T R (X_i - X_i^{\text{corr}})\right) \tag{5}$$

Set the rebuilt chaotic time series to x_i' using the inverse procedure of the Takens technique to reconstruct the chaotic time series. Choose the optimal wavelet basis function and decomposition layer for x_i' wavelet decomposition. The following are the coefficients of decomposition:

$$\begin{cases} C_{j+1}(k) = \sum_n C_j(n) p(n-2k) \\ D_{j+1}(k) = \sum_n D_j(n) q(n-2k) \end{cases} \tag{6}$$

where $C_{j+1}(k)$ represents the approximation coefficient and $D_{j+1}(k)$ represents the detail coefficient; p and q represent the impulse responses of the filter; and j represents the decomposition scale. The approximation wavelet coefficients are invariant, and after passing the soft threshold setting on the detail coefficients, $\widehat{x_i'}$ is rebuilt to yield the noise-reduced chaotic time series.

$$\widehat{x_i'} = \sum_n C_{j+1}(n) p(n-2k) + \sum_n D_{j+1}(n) q(n-2k) \tag{7}$$

B. Principle of Particle Swarm Optimization

A particle swarm algorithm is a kind of meta-heuristic algorithm subjected to a model of bird flock behavior. The birds in the flock are the particles in the particle swarm, and each particle has two attributes: velocity and position. Suppose the position of the i-th particle in the m-dimensional space is $K_i = (k_{i1}, k_{i2}, \ldots, k_{im})$, the velocity is $V_i = (v_{i1}, v_{i2}, \ldots, v_{im})$, and the current best position is $P_i = (p_{i1}, p_{i2}, \ldots, p_{im})$. The best position in all particles is $P_g = (p_{g1}, p_{g2}, \ldots, p_{gm})$. Each current extreme value P_{best} is shared with other particles in order to find the optimal solution G_{best} of the whole and to adjust its position and velocity, where the x+1th iteration is as follows:

$$v_{im}^{x+1} = w v_{im}^{x} + c_1 \varepsilon (p_{im}^{x} - k_{im}^{x}) + c_2 \eta (p_{gm}^{x} - k_{im}^{x}) \tag{8}$$

$$k_{im}^{x+1} = k_{im}^{x} + v_{im}^{x+1} \tag{9}$$

where w is the weighting factor controlling the flight speed; ε, η are random values obeying U(0,1) uniform distribution; c_1 and c_2 are learning factors used to adjust the weighting affecting the flight direction.

C. LW-PSO Algorithm

The local projection technique is the standard approach for denoising chaotic time series. It reconstructs the flow structure of a chaotic attractor orbit after phase space and splits the signal subspace and noise subspace to generate the noisy chaotic time series. Although it can reconstruct the original dynamical system to the greatest degree possible, its high frequency signal noise denoising capability is limited. In the detail coefficient portion of the local projection approach, high-frequency denoising is performed by defining a soft threshold for the system by wavelet transform. However, regardless of the denoising method employed, the original chaotic time series information will be lost, thus the lost information is retrieved by defining the goal function, combining noise reduction signals linearly, and optimising coefficients using the particle swarm algorithm. To preprocess the chaotic time series, a model based on the fusion of the local projection method, the wavelet transform, and particle swarm optimization (referred to as LW-PSO) is developed, which effectively reduces noise interference and preserves the original dynamical system to the greatest extent possible.

The specific steps of the LW-PSO algorithm are as follows:

1) Let x_i be the noisy chaotic time series of the reaction chaotic system, where the clean series is y_i.
2) The local projection method is used to process x_i to get x_i', Referring to equation 5.

$$x_i' = \widehat{X}_i \tag{10}$$

3) Using wavelet transform x_i', we obtain its approximation coefficients $C_{j+1}(k)$ and detail coefficients $D_{j+1}(k)$
4) Retain the approximate coefficients, soft threshold the detail coefficients, filter out the detail coefficients, and reconstruct the signal, Referring to equation 7.
5) Establish the objective function, find the minimum value.

$$f(x)=\frac{1}{n}\sum\left(y_i-a\times x_i-\mathrm{b}\times x_i'-\mathrm{c}\times x_i'\right)^2 \quad (11)$$

6) Particle swarm parameters initialization: given particle swarm size i, learning factors c_1 and c_2, maximum number of iterations ItTimes, variable dimension m, velocity size and position variation.
7) Calculate the fitness of each particle target and find the global optimal position p_{gm}^{x} and the optimal solution G_{best}.
8) Given the random parameter matrix ε, η, the inertia weight coefficient w varies linearly to update the particle velocity position.
9) Determine whether the preset number of iterations or the target is reached; if yes, stop the calculation and output p_{gm}^{x} and the optimal solution G_{best}, otherwise turn 7.

D. LW-PSO algorithm flow chart

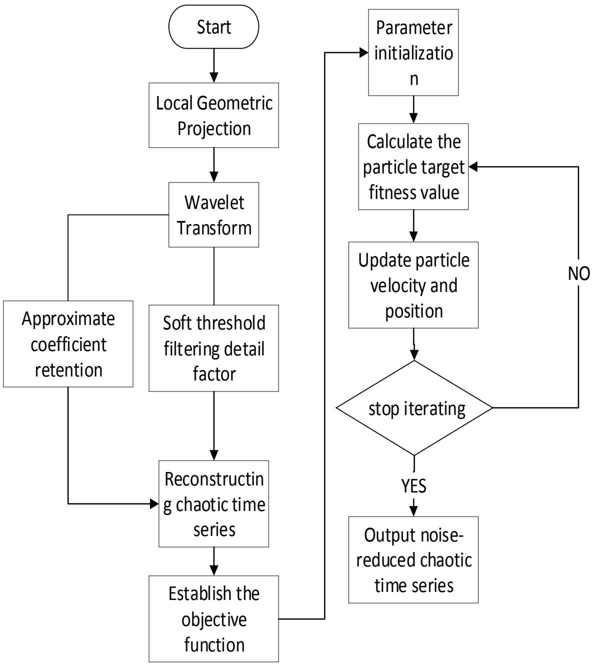

Start
Local Geometric Projection
Wavelet Transform
Approximate coefficient retention
Soft threshold filtering detail factor
Reconstructing chaotic time series
Establish the objective function
Parameter initialization
Calculate the particle target fitness value
Update particle velocity and position
stop iterating
NO
YES
Output noise-reduced chaotic time series

Fig. 1 Flow chart of algorithm.

III. Simulation Verification

In order to verify the LW-PSO method proposed in this paper, the method is applied to the classical Lorenz system simulation and analyzed as follows: Firstly, Gaussian white noise is added to the Lorenz chaotic system, and the noise reduction effects of the four methods are compared with local projection noise reduction, wavelet transform, Schreiber noise reduction, and the LW-PSO method, respectively, and the chaos recognition rate of this system under the influence of 30%(SNR=10dB) Gaussian white noise is also compared.

The dynamical equations of the Lorenz system:

$$\begin{cases}\dot{x}=-a(x-y)\\ \dot{y}=cx-y-xz\\ z=-bz+xy\end{cases} \quad (12)$$

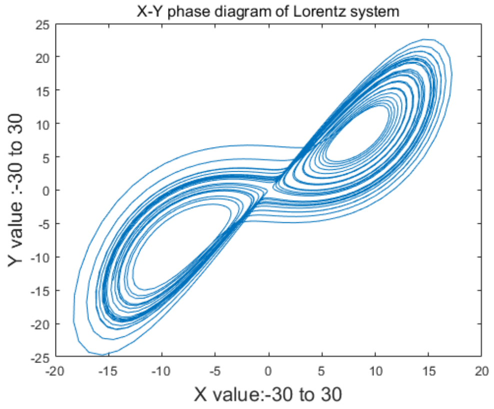


Fig. 2 Two-dimensional phase diagram of a noiseless Lorentzian system.

At a = 10, b = 8/3, c = 28, the system is in a chaotic state. Set the initial values $x_0 = -1, y_0 = 1, z_0 = 0$, Gaussian white noise level = 30% (or SNR=10 dB), constituting a chaotic system. Using four methods for noise reduction in this paper, Schreiber with the local projection noise reduction method reconstructs the phase space embedding dimension 7 with a delay time of 1. The wavelet function is chosen at bior6.8, and the high frequency noise of D1, D2, and D3 is chosen to be filtered out, while the low frequency is retained. The parameters in the LW-PSO method in PSO are set: the population size is 50; the initial learning factor is 2.0; the maximum number of iterations is 1000; and the variables are 2-dimensional. The results are shown in Fig. 3.

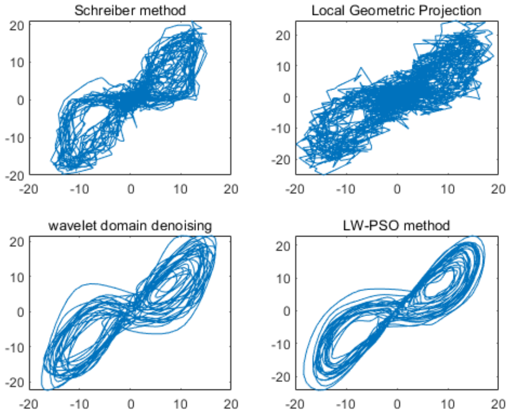


Fig. 3 As shown in the figure, the two-dimensional phase diagram after noise reduction of the four methods, from the figure can be known that the two-dimensional phase diagram after LW-PSO noise reduction is closer to the original phase diagram.

From Fig. 3, it can be seen that although the Schreiber noise reduction method, the local projection noise reduction method, and the wavelet transform method can reduce the noise to a certain extent, they cannot show the clear geometric structure of the attractor of the original Lorenz system. And the LW-PSO noise reduction method proposed in this paper can still recover the geometric mechanism of the original Lorenz system under

30% Gaussian white noise addition, restore the real dynamics trajectory of the signal, and the noise reduction effect is more superior to the other three methods. In order to analyze the effect specifically, the SNR, RMSE and chaos identification correct rate are calculated before and after the four methods.

As can be seen from Table I, the signal-to-noise ratios of the Schreiber denoising method, the local projection method, and the wavelet transform method after denoising are 13.18, 12.40, and 17.08, respectively. the chaos recognition rates are significantly improved to 82%, 58%, and 100% when adding 30% of Gaussian white noise intensity. The LW-PSO proposed in this paper has the best results in terms of S/N ratio, mean square error and chaos recognition rate: 16.24 for S/N ratio, 2.28 for mean square error reduction and 100% for chaos recognition rate, which are the best of the four methods.

TABLE I
COMPARISON OF THE RESULTS OF THE TWO METHODS

Comparison of the results of the four methods			
Three types of evaluation criteria			
System	***SNR***	***RMSE***	***CRR***
40% plus noise Lorenz system	10.00	2.70	0%
Schreiber noise reduction method	13.18	1.90	82%
Local Geometric Projection	12.40	2.08	58%
Wavelet domain denoising	17.08	1.21	100%
LW-PSO method	26.24	0.42	100%

To further validate the LW-PSO noise reduction method proposed in this paper, the method is applied to several classical systems: chaotic systems. Lorentzian system and Ikeda system; nonlinear stochastic system: sinusoidal system driven by noise. Meanwhile, 0%, 10%, 20%, 30% and 40% white noise are superimposed on the classical system with 5000 data volumes to observe whether the chaos recognition rate is improved after the noise reduction of the original method and the noise reduction of the improved method, and the experimental results are shown in Table II below.

TABLE II
CLASSICAL SYSTEMS IN VARYING DEGREES OF WHITE NOISE INFLUENCE

Correct Chaos Recognition Rate after Schreiber Noise Reduction [16]					
Measure noise level (% of std. dev.)					
System	**0%**	**10%**	**20%**	**30%**	**40%**
Lorenz	100%	100%	97%	82%	36%
Noise-driven sine map	50%	3%	22%	5%	78%
Ikeda	100%	100%	100%	100%	14%
Chaos recognition correct rate after noise reduction by R-S method					
Measure noise level (% of std. dev.)					
System	**0%**	**10%**	**20%**	**30%**	**40%**
Lorenz	100%	100%	100%	100%	76%
Noise-driven sine map	35%	85%	80%	45%	40%
Ikeda	100%	100%	100%	100%	90%

The noise-driven sinusoidal mapping is originally a random system, but it is easily identified as a chaotic system in the noisy environment. As can be seen from Table II, the Schreiber method performs better in the noise-free environment and the high-noise environment; under the LW-PSO method noise reduction, the correct recognition rate of the method system is higher in the 10%-30% noise environment, and the two methods can be applied in different environments. 0% to 30% noise added classical chaotic systems Lorenz system and Ikeda system, Schreiber In the classical chaotic system Lorenz system and Ikeda system with 0% to 30% noise addition, the recognition rate of the Schreiber method after noise reduction is above 82%, and the recognition rate of the LW-PSO method after noise reduction is above 100%, indicating that the difference between the recognition rate of the chaos after noise reduction of the two methods is not significant at low noise content and high signal-to-noise ratio. It is difficult to achieve the noise reduction effect only from the flow structure in the high noise environment, and the chaos recognition rate of the LW-PSO method is much higher than that of the Schreiber method when 40% noise is added.

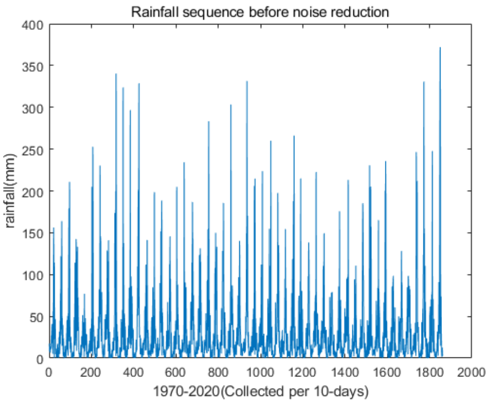
IV. Example Simulation

The chaotic nature of precipitation time series[17], [18]has been proved by Sivakumar et al. The observation of precipitation time series is susceptible to evaporation, loss, instrumentation, and human error in reading data to generate a lot of noise. The chaotic nature of precipitation time series is highly sensitive to noise, and the accuracy of precipitation time series data is very important for water development and river flood control, so effective noise reduction is needed for the rainfall time series.

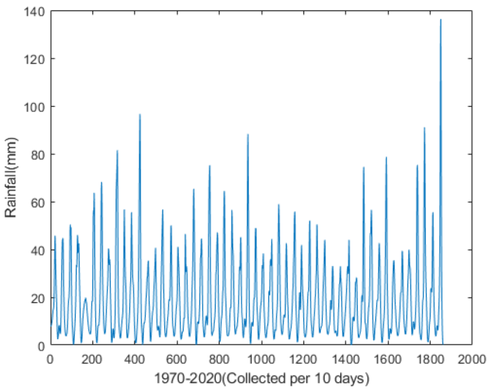
A. 50-year rainfall in Chengdu

The actual observed values of daily and decadal precipitation in Chengdu from January 1970 to December 2020 are shown in the figures. Noise reduction is applied to this data, and its noise reduction sequence is shown in Fig. 4.

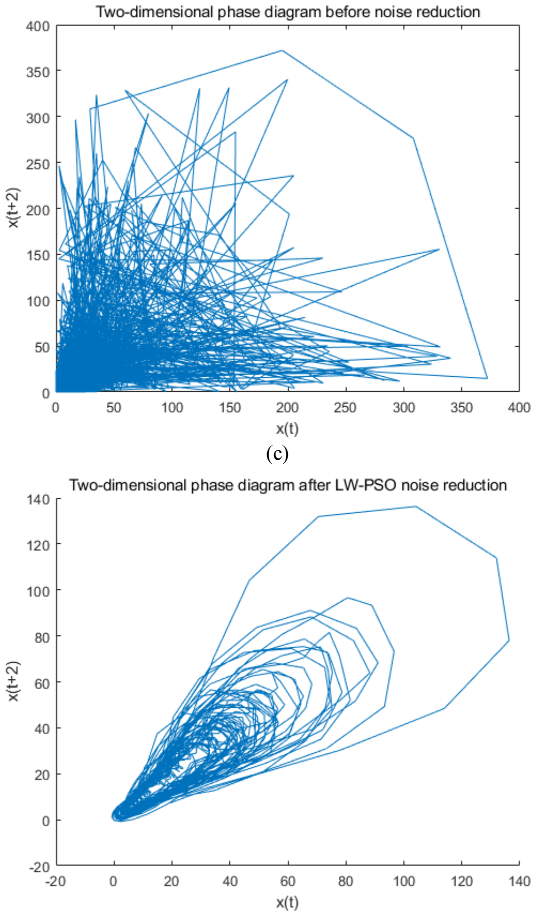
The rainfall is affected by noise, as shown in Fig. 4(a) (b), which causes the rainfall sequence map before noise reduction to appear messy, whereas the rainfall sequence after noise reduction is smooth and has some regularity.



(a)



(b)



(c)

(d)

Fig. 4 The actual observed values of precipitation per decade in Chengdu from January 1970 to December 2020, (a) shows the rainfall series with noise, (b) shows the rainfall series with LW-PSO method, (c) shows the phase diagram of the rainfall series with noise, and (d) shows the phase diagram of the rainfall series with LW-PSO method.

The rainfall is affected by noise, as shown in Fig. 4(a)(b), which causes the rainfall sequence map before noise reduction to appear messy, whereas the rainfall sequence after noise reduction is smooth and has some regularity. From the C-C method to find the embedding dimension m = 3 and the time delay τ = 2, it can be seen from the two-dimensional phase space reconstruction Fig. 4(c) that the phase diagram of the rainfall sequence before noise reduction shows the characteristics of randomness, while the two-dimensional phase Fig. 4 (d) of the rainfall sequence after noise reduction obviously shows the geometric structure of chaotic attractors, which indicates that the rainfall sequence retains its deterministic components and is obviously enhanced after noise reduction using the method in this paper. Not only is the noise reduction effect achieved, but also the chaotic attractor of the two-dimensional phase diagram once again verifies the existence of certain chaotic characteristics of the Chengdu rainfall sequence.

After the noise reduction by LW-PSO and the other three methods, the chaos identification of the rainfall sequences was carried out. The closer the value of the sequence result from the 0-1 test [19]is to 1, the stronger the chaos of the system, and the closer the value of the sequence result is to 0, the more chaotic the system is, as shown in Fig. 5. The daily precipitation for each five-year period from 1970 to 2020 is calculated as the rainfall series.

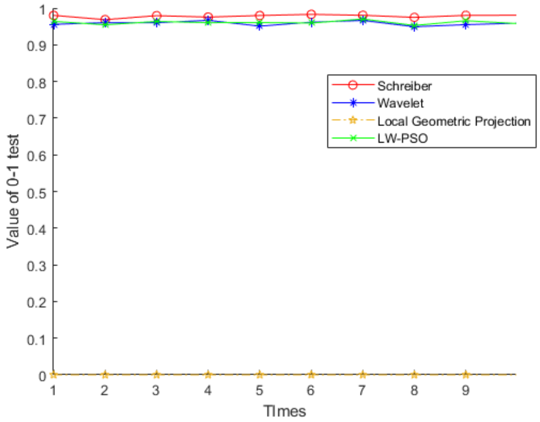


Fig. 5 The figure shows the K values of the sequence 0-1 test method after noise reduction by four methods.

Due to the influence of noise, the K-value of the Chengdu rainfall sequence is calculated to be close to 0 by the 0-1 test method before noise reduction, and its sequence exhibits random characteristics, which is consistent with our observation in the two-dimensional phase diagram. However, the K-value after noise reduction by the local projection method is also close to 0, indicating that the local projection method is less effective at low signal-to-noise ratios. The rainfall sequences after noise reduction by the Schreiber method, wavelet transform method, and LW-PSO method can all be tested for K-values and have similar K-values. It can be seen that although the 0-1 test K value of the local projection method is 0, the LW-PSO method proposed by this paper can not only solve the poor performance of the local projection method at low signal-to-noise ratio but also complete the recovery of the original rainfall sequence chaotic system under the particle swarm optimization algorithm. The K value of the 0-1 test method is close to 1, and the chaotic attractor in the two-dimensional phase diagram is more obvious. It indicates that the LW-PSO method filters out the Chengdu rainfall sequence noise more adequately and retains the original chaotic characteristics.

V. Conclusion

This paper proposes a chaotic sequence noise reduction method that integrates local projection method, wavelet change and particle swarm algorithm, and uses Lorenz, Ikeda and other classical chaotic sequences for simulation. Experimentally, it is proved that in the process of gradually increasing noise intensity, the original dynamical structure is destroyed more seriously, and it is difficult to achieve a better effect of noise reduction only from the perspective of phase space reconstruction embedded in the flow structure. And this paper starts from dividing the signal and noise subspace of the noisy chaotic sequence and removing the noise subspace; classifying the low frequency and high frequency by wavelet transform, removing the high frequency D1-D3, and retaining the original chaotic sequence signal to the maximum extent. The combined two noise reduction sequences are optimized and smoothed by particle swarm algorithm, and the signal-to-noise ratio, mean square error and chaos recognition rate are much better than the other three methods.

The simulation results show that the proposed LW-PSO method can be effectively applied to the noisy chaotic time series in Chengdu, which is ready for the next step of chaos prediction and chaos control.

REFERENCES

[1] Liu, X. X. , et al. "Noise reduction and prediction system of rainfall chaotic time series." Journal of Heb University of Engineering (2007).

[2] Banik, S. , et al. "Modeling chaotic behavior of Dhaka Stock Market Index values using the neuro-fuzzy model." International Conference on Computer & Information Technology IEEE, 2008.

[3] Ren, J. , Q. S. Zeng , and R. R. Wei . "Neural Network Forecasting Model for Sunspots Time Series Prediction Based on Phase Space Reconstruction." Computer Simulation (2014).

[4] Schreiber, T. , and H. Kantz . "Noise in chaotic data: Diagnosis and treatment." Chaos An Interdisciplinary Journal of Nonlinear Science 5.1(1995):133-142.

[5] Jaeger, L. , and H. Kantz . "Unbiased reconstruction of the dynamics underlying a noisy chaotic time series." Chaos 6.3(1996):440-450.

[6] Xu, D. , and F Lu. "Modeling global vector fields of chaotic systems from noisy time series with the aid of structure-selection techniques." Chaos 16.4(2006):450-458.

[7] Grassberger, P. , et al. "On noise reduction methods for chaotic data." Chaos 3.2(1993):127-141.

[8] Schreiber, and Thomas. "Extremely simple nonlinear noise-reduction method." Physical Review E Statistical Physics Plasmas Fluids & Related Interdisciplinary Topics 47.4(1993):2401.

[9] Leontitsis, A. , T. Bountis , and J. Pagge . "An adaptive way for improving noise reduction using local geometric projection." Chaos: An Interdisciplinary Journal of Nonlinear Science 14.1(2004):106-110.

[10] Mera, M. E. , and M. Mora?N . "Geometric noise reduction for multivariate time series." Chaos 16.1(2006):127-141.

[11] J. Gao, H. Sultan, J. Hu and W. Tung, "Denoising Nonlinear Time Series by Adaptive Filtering and Wavelet Shrinkage: A Comparison," in IEEE Signal Processing Letters, vol. 17, no. 3, pp. 237-240, March 2010, doi: 10.1109/LSP.2009.2037773.

[12] David, and Chelidze. "Smooth local subspace projection for nonlinear noise reduction. " Chaos (Woodbury, N.Y.) (2014).

[13] Moore, J. M. , M. Small , and A. Karrech . "Improvements to local projective noise reduction through higher order and multiscale refinements." Chaos 25.6(2015):653-671.

[14] Zhenglong W U , Wang Q , Liu K . Training neural networks with PSO for identification of chaotic systems[J]. Computer Engineering & Applications, 2008, 44(14):76-79.

[15] Yajing W U , Jun M A . De-Noising for Chaotic Signal Using PSO and Lifting Wavelet Transform[J]. Chinese Journal of Electron Devices, 2014.

[16] Toker, D. , F. T. Sommer , and M. D'Esposito . "A simple method for detecting chaos in nature." Communications Biology (2020).

[17] Sivakumar, B. , et al. "Singapore Rainfall Behavior: Chaotic?." Journal of Hydrologic Engineering 4.1(1999):38-48.

[18] Dhanya, C. T. , and D. N. Kumar . "Nonlinear ensemble prediction of chaotic daily rainfall." Advances in Water Resources 33.3(2010):327-347.

[19] Gottwald, G. A. , and I. Melbourne . "On the Validity of the 0-1 Test for Chaos." Nonlinearity 22.6(2009):1367-1382.

A Family of Distributions on Learnable Problems Without Uniform Convergence

César Garza

Abstract—In supervised binary classification and regression problems, it is well-known that learnability is equivalent to uniform convergence of the hypothesis class, and if a problem is learnable, it is learnable by empirical risk minimization. For the general learning setting of unsupervised learning tasks, there are non-trivial learning problems where uniform convergence does not hold. We present here the task of learning centers of mass with an extra feature that "activates" some of the coordinates over the unit ball in a Hilbert space. We show that the learning problem is learnable under a stable RLM rule. We introduce a family of distributions over the domain space with some mild restrictions for which the sample complexity of uniform convergence for these problems must grow logarithmically with the dimension of the Hilbert space. If we take this dimension to infinity, we obtain a learnable problem for which the uniform convergence property fails for a vast family of distributions.

Keywords—statistical learning theory, learnability, uniform convergence, stability, regularized loss minimization

I. Introduction

In papers such as Blumer et. al. [1], different conditions have been shown to be equivalent to learning supervised tasks such as binary classification, regression, or multiclass prediction. Learning with the empirical risk minimization rule (ERM) in the supervised case is equivalent to uniform convergence of the empirical risk to the true risk with a rate that is independent of the distribution over the instance set. For binary classification, Vapnik and Chervonenkis [2] showed that finiteness of a combinatorial condition known as the VC-dimension is a necessary and sufficient condition for learnability under the ERM rule. For some regression problems, finite fat-shattering dimension characterizes learnability [3] and the Natarajan dimension characterizes learnability of some multiclass learning problems [4].

For the general learning setting, there is no equivalence between learnability and uniform convergence, as Shalev-Shwartz et al. showed in [5]. Instead, the key notion is stability, as defined in section III. Examples of learnable problems without uniform convergence can shed more light into how Vapnik's notion of "strict" learnability fails in the framework of unsupervised learning.

In this paper we present two unsupervised tasks for the center of mass over the unit ball in some Hilbert spaces. The second problem is a modification of the first that makes it a strictly convex, bounded, smooth problem. We show that these tasks are learnable by exhibiting stability using smoothness of the corresponding loss functions, with a coefficient that is independent of the dimension d of the Hilbert space. We also show that these tasks possess distributions $\mathcal{D}$ concentrated in a small ball around the origin for which the uniform convergence property does not hold in the infinite dimensional case. In section II we present the formal definitions of learning under the ERM rule in the supervised case and discuss the equivalent notions of uniform convergence and finiteness of VC-dimension in the binary classification case. In section III we introduce the generalized concept of learning as defined in [5]. After defining the equivalent concept of stability, we present a theorem from [5] that shows that for convex-smooth-bounded problems, the Regularized Loss Minimization rule (RLM) with Tikhonov regularization leads to a stable learning algorithm. Finally, in section IV, we introduce our learning problem where the uniform convergence property fails. This can be described as the task of finding the "center of mass" of a distribution over the unit ball of $\mathbb{R}^d$, where an extra parameter $\boldsymbol{\alpha}$ indicates which of the coordinates are marked as "active" or "inactive". We show that this problem is learnable using the RLM rule with a sample complexity that does not depend on d. Then we choose a probability distribution for the instance space such that if $m < \log_2(d)$, there is a high probability that a sample of i.i.d. labeled points of size m has a high estimation error. We say that such samples are not "ϵ-representative". The main theorems of this paper are Theorems 5 and 6 where we show that for distributions $\mathcal{D}$ on the domain $\mathcal{Z}$ concentrated in a ball of radius $1/4$ and yielding a uniform Bernoulli distribution on the parameter $\boldsymbol{\alpha}$, a.s. the ERM rule does not converge to a minimizer of the true population risk as the sample size $m \to \infty$, not even for strictly convex bounded smooth problems where the ERM minimizer is unique. We hope that the distributions presented here are only the starting point for a rich variety of stable problems in unsupervised settings that lack the uniform convergence property.

II. The Supervised Learning Setting

In the supervised learning setting, we have an instant space $\mathcal{X}$, a label set $\mathcal{Y}$, and a hypothesis class $\mathcal{H}$. The domain $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$ has a sigma-algebra structure and we have a "loss" function $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}_{\geq 0}$ that is measurable for all $h \in \mathcal{H}$. We also assume the loss function is bounded over $\mathcal{H} \times \mathcal{Z}$.

Given a probability distribution $\mathcal{D}$ over $\mathcal{Z}$, the *risk* or *true error* of a hypothesis $h \in \mathcal{H}$ denoted as $L_{\mathcal{D}}(h)$ is defined as the expected value of the loss function over $\mathcal{Z}$; that is,

$$L_{\mathcal{D}}(h) = \underset{(\mathbf{x},y)\sim\mathcal{D}}{\mathbb{E}}[\ell(h,(\mathbf{x},y))]$$

While $\mathcal{H}, \mathcal{Z}$ and the loss function ℓ are known to the learner, we assume that $\mathcal{D}$ is unknown. It is thus not possible to simply choose $h \in \mathcal{H}$ that minimizes $L_{\mathcal{D}}(h)$. Instead, we consider

C. Garza is with the Department of Mathematics & Statistics, University of Houston Downtown, Houston, TX, 77002 USA e-mail: (garzace@uhd.edu).

training samples $S \sim \mathcal{D}^m$ of m i.i.d. draws from $\mathcal{Z}$. Each sample S is a sequence of the form $((\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m))$, where $\mathbf{x}_i \in \mathcal{X}$ and y_i is the corresponding label. For any $m \in \mathbb{N}$, we will use the notation $[m]$ to denote $\{1, \dots, m\}$. Our overall goal in this setting is to have a learning algorithm A that picks a hypothesis $A(S) \in \mathcal{H}$ based on the training sample S with approximately minimal possible risk. Generally, we expect the approximation to get better with the sample size. Before we give the formal definition of learnability, we present some examples of supervised statistical learning tasks:

- **Binary Classification:** Let $\mathcal{Y} = \{0, 1\}$ and let $\mathcal{H}$ be the set of functions $h : \mathcal{X} \to \{0, 1\}$. The loss function is the indicator function $\ell(h, (\mathbf{x}, y)) = \mathbb{1}_{h(\mathbf{x}) \neq y}$. This is also known as the $0 - 1$ loss function, which measures if h labeled the example $(\mathbf{x}, y)$ properly or not.

- **Linear Regression:** Let $\mathcal{X}$ be a bounded subset of $\mathbb{R}^n$ and let $\mathcal{Y}$ be a bounded subset of $\mathbb{R}$. Let $\mathcal{H}$ be a set of bounded functions $h : \mathcal{X} \to \mathbb{R}$, and let ℓ be the square loss function: $\ell(h, (\mathbf{x}, y)) = (h(\mathbf{x}) - y)^2$.

- **Ranking:** We can consider ranking problems for classification or information retrieval purposes. The training data is a list of items and we assign a partial order to the items in the list. If $\mathcal{X}$ is the set of instances, let $\mathcal{X}^* = \bigcup_{n=1}^{\infty} \mathcal{X}^n$ be the set of all sequences of instances from $\mathcal{X}$ of arbitrary length. Here $\mathcal{Z} = \bigcup_{r=1}^{\infty} (\mathcal{X}^r \times \mathbb{R}^r)$. The hypothesis class $\mathcal{H}$ is the set of ranking hypotheses h that receive a sequence of instances $\overline{\mathbf{x}} = (\mathbf{x}_1, \dots, \mathbf{x}_r) \in \mathcal{X}^*$ and return a vector $\mathbf{y} \in \mathbb{R}^r$. By sorting the elements of $\mathbf{y}$ in increasing order, we obtain a permutation of $[r]$.
  There are many possible ways to define a loss function for ranking. If we denote by $\pi(\mathbf{y})$ the permutation of $[r]$ induced by the vector $\mathbf{y} \in \mathbb{R}^r$, then one example is the $0 - 1$ loss function $\ell(h, (\overline{\mathbf{x}}, \mathbf{y})) = \mathbb{1}_{[\pi(h(\overline{\mathbf{x}})) \neq \pi(\mathbf{y})]}$. Better examples of loss functions for ranking are the Kendall-Tau loss or the Normalized Discounted Cumulative Gain loss. See [6] for more details.

Ideally, we wish to pick in this setting a hypothesis $h \in \mathcal{H}$ that minimizes the true risk $L_{\mathcal{D}}(h)$, but since $\mathcal{D}$ is unknown to the learner, this is not feasible. We wish to obtain a learning rule A such that, upon receiving a training sample S of size m, A outputs a hypothesis $A(S)$ and the expected value of the difference between the true risk of $A(S)$ and the minimal risk is small, with this value approaching 0 as the sample size $m \to \infty$. That is,

$$\mathop{\mathbb{E}}_{S \sim \mathcal{D}^m} [L_{\mathcal{D}}(A(S)) - \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)] \leq \epsilon(m)$$

where $\mathcal{D}^m$ is the probability over m-tuples in $\mathcal{Z}$ induced by applying $\mathcal{D}$ to pick each element of the tuple independently of the other members of the tuple. We also require the rate $\epsilon(m)$ to be monotonically decreasing with $\epsilon(m) \xrightarrow{m \to \infty} 0$.

Since $\mathcal{D}$ is unknown, we ask for learnability that the above inequality is consistent over *all* distributions $\mathcal{D}$ on $\mathcal{Z}$. This leads us to the formal definition of learnability of supervised tasks.

Definition 1. A learning problem is *learnable* if there exist a learning rule A and a monotonically decreasing sequence $\epsilon_{\text{const}}(m)$, such that $\epsilon_{\text{const}}(m) \xrightarrow{m \to \infty} 0$ and for all distributions $\mathcal{D}$ on $\mathcal{Z}$,

$$\mathop{\mathbb{E}}_{S \sim \mathcal{D}^m} [L_{\mathcal{D}}(A(S)) - \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)] \leq \epsilon_{\text{const}}(m). \tag{1}$$

A learning rule A for which this holds is denoted as a *universally consistent* learning rule.

This is a direct generalization of agnostic PAC-learnability as seen in [3]. Note that instead of asking for an inequality similar to (1) that holds with probability $1 - \delta$ over all samples S, we ask for a uniform rate over the expected value of the difference of errors for all distributions on $\mathcal{Z}$.

A. Equivalent Forms of Learnability

The learner does not have access to the distribution $\mathcal{D}$ of the domain. Nevertheless, the learner can compute an *empirical error* or *empirical risk* based on the training sample S. This is denoted by $L_S(h)$ and it is defined as the error a hypothesis h incurs over the training sample. If $S = ((\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m))$, then

$$L_S(h) := \frac{1}{m} \sum_{i=1}^{m} \ell(h, (\mathbf{x}_i, y_i))$$

We say that a rule A is an ERM (*Empirical Risk Minimizer*) if it minimizes the empirical risk

$$A(S) \in \operatorname*{arg\,min}_{h \in \mathcal{H}} L_S(h).$$

Here argmin denotes the collection of hypotheses in $\mathcal{H}$ for which the value of $L_S(h)$ over $\mathcal{H}$ is minimal.

We say that a problem is learnable under the ERM rule if the ERM rule described above satisfies (1) for all distributions $\mathcal{D}$ over $\mathcal{Z}$.

A simple idea that is related to learnability is to have a hypothesis class $\mathcal{H}$ for which the empirical risk of any hypothesis $h \in \mathcal{H}$ is a good approximation of its true risk. This is formalized in the definition of *uniform convergence* of a learning problem.

Definition 2. A learning problem with domain $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and hypothesis class $\mathcal{H}$ is said to have the *uniform convergence property* if

$$\sup_{\mathcal{D}} \mathop{\mathbb{E}}_{S \sim \mathcal{D}^m} \left[\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \right] \xrightarrow{m \to \infty} 0$$

More intuitively, given any $\epsilon > 0$, there exists $m \in \mathbb{N}$ such that for any distribution $\mathcal{D}$ on $\mathcal{Z}$ and any hypothesis $h \in \mathcal{H}$, the mean value of $|L_{\mathcal{D}}(h) - L_S(h)|$ is less than ϵ.

The uniform convergence property says that the empirical risks of hypotheses in the hypothesis class converges to their population risk uniformly, with a distribution-independent rate.

We offer a third combinatorial concept that is used in binary classification problems only. Let $\mathcal{Z} = \mathcal{X} \times \{0, 1\}$, where each hypothesis $h \in \mathcal{H}$ is a mapping $h : \mathcal{X} \to \{0, 1\}$ and ℓ is the $0 - 1$ loss function $\ell(h, (\mathbf{x}, y)) = \mathbb{1}_{h(\mathbf{x}) \neq y}$.

Definition 3. Let C be a finite subset of $\mathcal{X}$. We say that a hypothesis class $\mathcal{H}$ *shatters* C if any function from C to $\{0,1\}$ can be obtained as a restriction of an element $h \in \mathcal{H}$ to C.

Vapnik and Chervonenkis defined in [2] a simple combinatorial measure that implies uniform convergence.

Definition 4. Let $\mathcal{H}$ be a hypothesis class. The VC-dimension of $\mathcal{H}$, denoted VCdim($\mathcal{H}$), is the maximal cardinal D such that a set of cardinality D in $\mathcal{X}$ is shattered by $\mathcal{H}$.

For binary classification problems we have a chain of equivalences explained in the next theorem.

Theorem 1 (The Fundamental Theorem of Statistical Learning, see [6] Theorem 6.7)**.** *Let $\mathcal{X}$ be the set of instances and let $\mathcal{H}$ be a hypothesis class of binary functions on $\mathcal{X}$. Then, under the $0-1$ loss function, the following are equivalent:*

1) *$\mathcal{H}$ has a finite VC-dimension.*
2) *$\mathcal{H}$ has the uniform convergence property.*
3) *Any ERM rule is a successful learner for $\mathcal{H}$.*
4) *$\mathcal{H}$ is learnable according to Definition 1.*

The situation is depicted in Figure 1.

In the case of regression problems, a similar characterization holds. This time a hypothesis h is a real-valued function $h : \mathcal{X} \to \mathbb{R}$ and the loss function is the squared-loss function $\ell(h,(\mathbf{x},y)) = (h(\mathbf{x}) - y)^2$. The VC dimension is replaced by the fat-shattering dimension, but the basic equivalence still holds: a problem is learnable if and only if uniform convergence holds if and only if the uniform convergence property is present (see [7]).

Remark 1. In our definitions of learnability, uniform convergence, and stability in the next section, we have used convergence in expectation, and defined the rates as rates on the expectation. Since the loss function ℓ is bounded, by the dominated convergence theorem, convergence in expectation is equivalent to convergence in probability. Furthermore, using Markov's inequality we can translate a rate of the form $\mathbb{E}[|X|] \leq \epsilon(m)$ to a "low confidence" guarantee $\mathbb{P}[|X| > \epsilon(m)/\delta] \leq \delta$. Thus "learnability" can be replaced with agnostic PAC learnability as defined in [6] in Theorem 1. For simplicity, we will not discuss in this paper the computational aspects of learnability, although for the tasks presented here there are well-known efficient algorithms such as SGD that solve the problem.

III. General Learning Framework

We now consider the general learning setting, where the domain $\mathcal{Z}$ is an arbitrary measurable space. There is still a hypothesis class $\mathcal{H}$ and a loss function $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}_{\geq 0}$ that is measurable on $\mathcal{Z}$ and bounded by some constant B. That is, $\ell(h,\mathbf{z}) \leq B$ for all $h \in \mathcal{H}$ and $\mathbf{z} \in \mathcal{Z}$.

Some examples of general learning tasks that do not fit in the supervised setting are:

- **K-means clustering:** Let $\mathcal{Z} = \mathbb{R}^n$, let $\mathcal{H}$ be all subsets of $\mathbb{R}^n$ with k elements, and let $\ell(h,\mathbf{z}) = \min_{\mathbf{c}\in h} \|\mathbf{c} - \mathbf{z}\|^2$. Here, each h represents a set of k centroids, and ℓ measures the square of the Euclidean distance between an instance $\mathbf{z}$ and its nearest centroid, according to the hypothesis h.
- **Stochastic Convex Optimization in Hilbert Spaces:** Let $\mathcal{Z}$ be any measurable set, let $\mathcal{H}$ be a closed, convex and bounded subset of a Hilbert space, and let $\ell(h,\mathbf{z})$ be Lipschitz and and convex with respect to its first argument. The task is to minimize the true risk function $L_{\mathcal{D}}(h) = \mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[\ell(h,\mathbf{z})]$, where the distribution $\mathcal{D}$ over Z is unknown, based on a training sample $S = (\mathbf{z}_1,\ldots,\mathbf{z}_m)$.

The definitions of learnability and uniform convergence in the general case are exactly as in Definitions 1 and 2 respectively, with the only difference being a more general domain space $\mathcal{Z}$.

In the next section we will prove that in the general framework learnability is no longer equivalent to uniform convergence. We will define an equivalent notion of learnability that is no longer concerned about the complexity of the hypothesis class. Instead, we wish to control the variance of the learning rule. Intuitively, an algorithm is considered stable if a slight change of its input does not change its output much. To be more precise, given the training set S and an additional example $\mathbf{z}'$ from $\mathcal{Z}$, let $S^{(i)}$ be the training set obtained by replacing the i'th example of S with $\mathbf{z}'$. That is,

$$S^{(i)} = (\mathbf{z}_1,\ldots,\mathbf{z}_{i-1},\mathbf{z}',\mathbf{z}_{i+1},\ldots,\mathbf{z}_m)$$

By "a small change of the input" we mean that we feed the learner A the sample $S^{(i)}$ instead of S. Observe that only one training sample is replaced. We then compare the loss of the hypothesis $A(S)$ on the element $\mathbf{z}_i$ to the loss of $A(S^{(i)})$ on the same element $\mathbf{z}_i$. We say that A is a *stable* algorithm if changing a single example in the training set does not lead to a significant change. Formally,

Definition 5. Let $\epsilon_{\text{st}}(m)$ be a monotonically decreasing function with $\epsilon_{\text{st}}(m) \xrightarrow{m\to\infty} 0$ and let $U(m)$ be the uniform distribution over $[m]$. We say that a learning algorithm A is on-average-replace-one-stable with rate $\epsilon_{\text{st}}(m)$ if for every distribution $\mathcal{D}$ over $\mathcal{Z}$

$$\mathop{\mathbb{E}}_{(S,\mathbf{z}')\sim\mathcal{D}^{m+1},i\sim U(m)}\left[\ell(A(S^{(i)}),\mathbf{z}_i) - \ell(A(S),\mathbf{z}_i)\right] \leq \epsilon_{\text{st}}(m)$$

For simplicity, we will call a learning algorithm that is on-average-replace-one-stable just *universally stable* or simply *stable*.

For supervised learning tasks like binary classification or regression, by the Fundamental Theorem of Statistical Learning, if a problem is learnable then it is learnable under any ERM rule. This is no longer true in the general setting. In this case, the correct approach is to choose a rule that is "asymptotically" ERM or AERM for short. The precise definition is as follows.

Definition 6. A rule A is *universally* an AERM rule with rate $\epsilon_{\text{erm}}(m) \xrightarrow{m\to\infty} 0$ if

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[L_S(A(S)) - \min_{h\in\mathcal{H}} L_S(h)] \leq \epsilon_{\text{erm}}(m)$$

for all distributions $\mathcal{D}$ over $\mathcal{Z}$.

In [5], Shalev-Shwarz et al. proved that general learnability is equivalent to having a stable universal AERM rule.

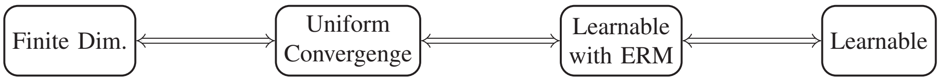


Fig. 1. The Fundamental Theorem of Statistical Learning

Theorem 2 ([5], Theorem 7)**.** *A learning problem is learnable if and only if there exists a stable universally AERM learning rule.*

The theorem even relates the three distinct convergence rates $\epsilon_{\text{const}}(m), \epsilon_{\text{erm}}(m)$, and $\epsilon_{\text{st}}(m)$, although we shall not be concerned about this. It should also be mentioned that, contrary to binary classification, not any AERM rule is enough for learnability; the AERM rule must also be stable. Figure 2 illustrates the correspondences in the general learning framework.

Note that the uniform convergence property no longer appears as an equivalent condition for learnability. It can be shown (see [5]) that uniform convergence is a sufficient condition for a stable universally AERM rule and hence learnability, but it is by no means a necessary condition.

In the next section, we consider two types of learning problems that are learnable by a stable universally AERM rule, but do not possess the uniform convergence property.

IV. A Learnable Problem without Uniform Convergence

Let $\mathcal{B}$ be the unit ball in $\mathbb{R}^d$, let $\mathcal{H} = \mathcal{B}$, and let $\mathcal{Z} = \mathcal{B} \times [0,1]^d$. Define a loss function $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}_{\geq 0}$ as follows:

$$\ell(\mathbf{h}, (\mathbf{x}, \boldsymbol{\alpha})) = \sum_{i=1}^{d} \alpha_i (x_i - h_i)^2 = \left\| \sqrt{\boldsymbol{\alpha}} * (\mathbf{x} - \mathbf{h}) \right\|^2 \quad (2)$$

where $\sqrt{\boldsymbol{\alpha}}$ is the element-wise square root and $\mathbf{u} * \mathbf{v}$ denotes an element-wise product. This is an unsupervised learning task where we try to find the "center of mass" of the distribution over $\mathcal{B}$ and the vector $\boldsymbol{\alpha}$ represents a vector of stochastic per-coordinate "confidence" weights α_i for each coordinate in $\mathbb{R}^d$.

We will prove that this problem is learnable using smoothness properties of the loss function. First we define formally the concept of smoothness that we will use in this paper.

Definition 7. A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is β-smooth if its gradient is β-Lipschitz. That is, for all $\mathbf{v}, \mathbf{w}$ in $\mathbb{R}^d$ we have $\|\nabla f(\mathbf{v}) - \nabla f(\mathbf{w})\| \leq \beta \|\mathbf{v} - \mathbf{w}\|$.

Now we can show that our loss function is β-smooth in its first argument for a constant β that does not depend on the dimension d.

Lemma 1. *The loss function $\ell(\cdot, (\mathbf{x}, \boldsymbol{\alpha}))$ in* (2) *is* 2*-smooth for all $d \in \mathbb{N}$.*

Proof 1. *Fix $(\mathbf{x}, \boldsymbol{\alpha})$ in $\mathcal{Z}$. Then for $\mathbf{v}, \mathbf{w}$ in $\mathcal{H}$,*

$$\begin{aligned} \|\nabla \ell(\mathbf{v}) - \nabla \ell(\mathbf{w})\| &= 2 \|\langle \alpha_1 (v_1 - w_1), \ldots, \alpha_d (v_d - w_d) \rangle\| \\ &\leq 2 \|\mathbf{v} - \mathbf{w}\| \end{aligned}$$

where the last inequality follows since each α_i satisfies $0 \leq \alpha_i \leq 1$. □

We have thus a learning problem $(\mathcal{H}, \mathcal{Z}, \ell)$ where the following holds:

- $\mathcal{H}$ is a convex bounded subset of $\mathbb{R}^d$.
- For all $\mathbf{z} \in \mathcal{H}$, the loss function $\ell(\cdot, \mathbf{z})$ is a convex, nonnegative, 2-smooth function such that $\ell(\mathbf{0}, \mathbf{z}) = \sum_{i=1}^{d} \alpha_i x_i^2 \leq \|\mathbf{x}\|^2 \leq 1$.

This is known as a Convex-Smooth-Bounded Learning problem (see [6, Definition 12.13]). Instead of working directly with the loss function ℓ, we use a "regularized" version of it; namely, we use an ERM rule for the regularized loss function

$$\ell(\mathbf{h}, \mathbf{z}) + \frac{\lambda}{2} \|\mathbf{h}\|^2$$

for some parameter $\lambda > 0$ to be chosen later. This is also known as the "Regularized Loss Minimization" (RLM) rule. The extra function $\lambda \|\mathbf{h}\|^2 / 2$ is known as Tikhonov regularization. If A is an RLM learner for some parameter $\lambda > 0$, then upon receiving a sample $S \sim \mathcal{D}^m$, the algorithm returns a hypothesis

$$A(S) \in \operatorname*{arg\,min}_{\mathbf{h} \in \mathcal{H}} \left(\ell(\mathbf{h}, \mathbf{z}) + \frac{\lambda}{2} \|\mathbf{h}\|^2 \right)$$

Theorem 3 says that the RLM rule is a successful learner for Convex-Smooth-Bounded Learning problems with a suitable boundedness condition on the loss function.

Theorem 3 ([6], Corollary 13.11)**.** *Let $(\mathcal{H}, \mathcal{Z}, \ell)$ be a convex-smooth-bounded learning problem with parameters β, B, where $\|\mathbf{h}\| \leq B$ for all $\mathbf{h} \in \mathcal{H}$. Assume in addition that $\ell(\mathbf{0}, \mathbf{z}) \leq 1$ for all $\mathbf{z} \in \mathcal{Z}$. For any $\epsilon \in (0,1)$, let $m \geq \dfrac{150 \beta B^2}{\epsilon^2}$ and set $\lambda = \epsilon / (3B^2)$. Let A be an RLM learner with parameter λ. Then, for every distribution $\mathcal{D}$ of $\mathcal{Z}$,*

$$\mathop{\mathbb{E}}_{S \sim \mathcal{D}^m} \left[L_{\mathcal{D}}(A(S)) - \min_{\mathbf{h} \in \mathcal{H}} L_{\mathcal{H}}(\mathbf{h}) \right] \leq \epsilon$$

For our center of mass problem, $\beta = 2$ and $B = 1$. Thus the RLM rule is stable and the problem is learnable under Definition 1 for any $d \in \mathbb{N}$.

Now take $\mathcal{H}$ to be the unit sphere $\mathcal{B}$ of an infinite-dimensional Hilbert space with orthonormal basis $\mathbf{e}_1, \mathbf{e}_2, \ldots$, where for $\mathbf{v} \in \mathcal{H}$, we refer to its coordinates $\mathbf{v}_j = \langle \mathbf{v}, \mathbf{e}_j \rangle$. The weights $\boldsymbol{\alpha}$ are now a mapping of each coordinate to $[0,1]$. That is, $\boldsymbol{\alpha}$ is an infinite sequence of reals in $[0,1]$. The loss function in (2) is defined with respect to this orthonormal basis and is still well-defined in this Hilbert space. Since $\beta = 2$ was independent of the dimension d, the infinite-dimensional problem $(\mathcal{H}, \mathcal{Z}, \ell)$ is still a convex-smooth-bounded learning problem and we thus obtain

Theorem 4. *Let $(\mathcal{H}, \mathcal{Z}, \ell)$ be the infinite-dimensional problem where $\mathcal{H} = \mathcal{B}$, $\mathcal{Z}$ is formed of pairs $(\mathbf{x}, \boldsymbol{\alpha})$ where $\mathbf{x} \in \mathcal{B}$ and*

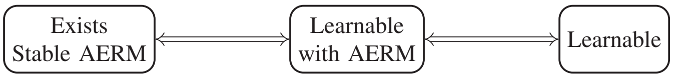


Fig. 2. The General Learning Framework

$\boldsymbol{\alpha}$ is a sequence of numbers in $[0,1]$, and $\ell(\mathbf{h},(\mathbf{x},\boldsymbol{\alpha}))$ is as in (2)*. Then $(\mathcal{H},\mathcal{Z},\ell)$ is a convex-smooth bounded problem learnable under a stable RLM rule.*

Next we present a family of distributions on $\mathcal{Z}$ for which $\sup_{\mathbf{h}\in\mathcal{H}}|L_{\mathcal{D}}(\mathbf{h})-L_S(\mathbf{h})|$ does not converge in mean to 0 as $m\to\infty$, showing that the uniform convergence property fails for this problem. This is an extension of the work in [5], where only one such distribution was showed. The main goal of this paper is to show how easy it is to find distributions where the true risk of a hypothesis is considerably bigger than the empirical risk, even when the true risk converges in mean to the minimal risk achievable by elements in the hypothesis class.

We start with the finite-dimensional case. Let d be a positive integer and consider the learning problem $(\mathcal{H},\mathcal{Z},\ell)$ defined above where $\mathcal{H}$ is the unit ball $\mathcal{B}$ in $\mathbb{R}^d$. Let $\mathcal{D}$ be a distribution only over $\mathcal{B}\times\{0,1\}^d$ satisfying the following two conditions:

1) $\underset{(\mathbf{x},\boldsymbol{\alpha})\sim\mathcal{D}}{\mathbb{P}}(\|\mathbf{x}\|>1/4)=0$
2) $\forall i\in[d]:\underset{(\mathbf{x},\boldsymbol{\alpha})\sim\mathcal{D}}{\mathbb{P}}(\boldsymbol{\alpha}_i=1)=\dfrac{1}{2}$

The two conditions state that the distribution is 0 away from the ball of radius $1/4$ in $\mathbb{R}^d$ and that the marginal distribution on $\{0,1\}^d$ is a sum of independent, uniform Bernoulli random variables. Here are two examples of such distributions:

Let C be a finite subset of $\mathcal{B}$ such that $\|\mathbf{x}\|\leq 1/4$ for all $\mathbf{x}\in C$. Let $\mathcal{D}_1$ be the uniform distribution on $C\times\{0,1\}^d$. More generally, let $C=\{\mathbf{x}_1,\mathbf{x}_2,\ldots\}$ be a denumerable collection in $\mathbb{R}^d$ such that $\|\mathbf{x}_i\|\leq 1/4$ for all i. Assign to each $(\mathbf{x}_i,\boldsymbol{\alpha})$ a probability of 2^{-i-d}. This yields a distribution $\mathcal{D}_2$ satisfying the two conditions above.

As a generalization of the previous example, let μ be any probability measure on $\mathcal{B}_{1/4}$, the ball of radius $1/4$ in $\mathbb{R}^d$ centered at $\mathbf{0}$ and let U be the uniform distribution on $\{0,1\}^d$. Then for $\mathcal{D}_3=\mu\times U$ extended to 0 over $\mathcal{B}\times\{0,1\}^d$ the two conditions hold.

We will show that the rate of uniform convergence for the problem $(\mathcal{H},\mathcal{Z},\ell)$ grows with d. First we define a notion of “representative” samples with respect to a distribution $\mathcal{D}$.

Definition 8. Let $\epsilon>0$. A training set S is called ϵ-representative (with respect to domain $\mathcal{Z}$, hypothesis class $\mathcal{H}$, loss function ℓ, and distribution $\mathcal{D}$) if

$$\forall\mathbf{h}\in\mathcal{H},\qquad |L_{\mathcal{D}}(\mathbf{h})-L_S(\mathbf{h})|\leq\epsilon$$

Lemma 2. *Let $\mathcal{D}$ be a distribution over $\mathcal{B}\times\{0,1\}^d$ satisfying* (1) *and* (2)*. Assume $2^m<d$. Then with probability of at least $1-e^{-1}$, a sample S of size m is not $\frac{1}{5}$-representative w.r.t. $(\mathcal{H},\mathcal{Z},\ell,\mathcal{D})$.*

Proof 2. *Let $S=((\mathbf{x}^{(1)},\boldsymbol{\alpha}^{(1)}),\ldots,(\mathbf{x}^{(m)},\boldsymbol{\alpha}^{(m)}))$ be a sample of m i.i.d. draws from $\mathcal{Z}$ with distribution $\mathcal{D}$. We will show that with probability at least $1-e^{-1}>0.63$, there exists a coordinate $j\in[d]$ such that $\boldsymbol{\alpha}_j^{(i)}=0$ for all $i\in[m]$.*

Indeed, the probability that this occurs is given by

$$\mathbb{P}\left(\bigcup_{j\in[d]}\bigcap_{i\in[m]}\{\boldsymbol{\alpha}_j^{(i)}=0\}\right)=1-\mathbb{P}\left(\bigcap_{j\in[d]}\bigcup_{i\in[m]}\{\boldsymbol{\alpha}_j^{(i)}=1\}\right)$$

By our choice of $\mathcal{D}$, the $\boldsymbol{\alpha}_j^{(i)}$ are independent uniform Bernoulli random variables. Hence

$$\begin{aligned}1-\mathbb{P}&\left(\bigcap_{j\in[d]}\bigcup_{i\in[m]}\{\boldsymbol{\alpha}_j^{(i)}=1\}\right)\\&=1-\prod_{j\in[d]}\left(1-\mathbb{P}\left(\boldsymbol{\alpha}_j^{(1)}+\ldots+\boldsymbol{\alpha}_j^{(m)}=0\right)\right)\\&=1-(1-2^{-m})^d\\&\geq 1-(e^{-2^{-m}})^d\\&=1-e^{-d2^{-m}}\\&\geq 1-e^{-1}\end{aligned}\tag{3}$$

Now we show that a sample S for which $\boldsymbol{\alpha}_j^{(i)}=0$ for some coordinate $j\in[d]$ and all $i\in[m]$ cannot be $\frac{1}{5}$ representative with respect to this distribution $\mathcal{D}$. Let $\mathbf{e}_j$ be the standard unit vector along coordinate j in $\mathbb{R}^d$. Then $\mathbf{e}_j\in\mathcal{H}$ and

$$L_S(\mathbf{e}_j)=\frac{1}{m}\sum_{i\in[m]}\ell(\mathbf{e}_j,(\mathbf{x}^{(i)},\boldsymbol{\alpha}^{(i)}))$$

since $\boldsymbol{\alpha}_j^{(i)}=0$ for all i and $\mathbf{e}_j$ has only one nonzero coordinate,

$$\begin{aligned}&=\frac{1}{m}\sum_{i\in[m]}\sum_{k\in[d]\setminus\{j\}}\boldsymbol{\alpha}_k^{(i)}(\mathbf{x}_k^{(i)})^2\\&\leq\frac{1}{m}\sum_{i\in[m]}\left\|\mathbf{x}^{(i)}\right\|^2\\&\leq\frac{1}{16}\end{aligned}$$

On the other hand, by the law of total expectation,

$$\begin{aligned}L_{\mathcal{D}}(\mathbf{e}_j)&=\underset{(\mathbf{x},\boldsymbol{\alpha})\sim\mathcal{D}}{\mathbb{E}}[\ell(\mathbf{e}_j,(\mathbf{x},\boldsymbol{\alpha}))]\\&=\underset{(\mathbf{x},\boldsymbol{\alpha})}{\mathbb{E}}[\ell(\mathbf{e}_j,(\mathbf{x},\boldsymbol{\alpha}))|\boldsymbol{\alpha}_j=1]\,\mathbb{P}(\boldsymbol{\alpha}_j=1)\\&+\underset{(\mathbf{x},\boldsymbol{\alpha})}{\mathbb{E}}[\ell(\mathbf{e}_j,(\mathbf{x},\boldsymbol{\alpha}))|\boldsymbol{\alpha}_j=0]\,\mathbb{P}(\boldsymbol{\alpha}_j=0)\\&\geq\frac{1}{2}\underset{(\mathbf{x},\boldsymbol{\alpha})}{\mathbb{E}}[(\mathbf{x}_j-1)^2+\ldots]\\&\geq\frac{1}{2}\left(\frac{3}{4}\right)^2=\frac{9}{32}\end{aligned}$$

We conclude that with probability of at least $1-e^{-1}$ we obtain a sample S of size m for which $|L_{\mathcal{D}}(\mathbf{e}_j)-L_S(\mathbf{e}_j)|\geq$

$\frac{9}{32} - \frac{1}{16} > \frac{1}{5}$ *for some* $j \in [d]$. *Such a sample is not* $\frac{1}{5}$-*representative.* □

Lemma 2 shows that the sample complexity m for uniform convergence in the $(\mathcal{H}, \mathcal{Z}, \ell)$ finite-dimensional problem is $\Omega(\log(d))$.

In the infinite dimensional case where $\mathcal{H} = \mathcal{B}$ is the unit sphere in a Hilbert space with orthonormal basis $\mathbf{e}_1, \mathbf{e}_2, \ldots$ and ℓ has the coordinate-free form in (2), we consider distributions $\mathcal{D}$ of the following form:

1) $\mathcal{D}$ is nonzero only on $\mathcal{B}_{1/4} \times [0,1]$, where we have identified sequences $\boldsymbol{\alpha}$ in $\{0,1\}$ with real numbers in the interval $[0,1]$.
2) The marginal distribution on $[0,1]$ is the uniform distribution. If we regard $\boldsymbol{\alpha} \in [0,1]$ as a sequence where each $\boldsymbol{\alpha}_j$ is 0 or 1, this means that all $\boldsymbol{\alpha}_j$ are independent uniform Bernoulli random variables.

Theorem 5. *Let* $(\mathcal{H}, \mathcal{Z}, \ell)$ *be the learnable problem as in Theorem 4. Then the problem does not have the uniform convergence property.*

Proof 5. *Let* $\mathcal{D}$ *be a distribution on* $\mathcal{Z}$ *with the two properties defined above. The estimates in* (3) *carry out the same in the infinite dimensional case. If we take the limit as* $d \to \infty$, *we obtain that a.s. if we take a training sample* S *of size* m, *there is a coordinate* j *such that* $\boldsymbol{\alpha}_j^{(i)} = 0$ *for all* $i \in [m]$. *By the same computations as in Lemma 2, we obtain that for such distributions* $\mathcal{D}$ *and for all* m,

$$\mathop{\mathbb{E}}_{S \sim \mathcal{D}^m}\left[\sup_{\mathbf{h} \in \mathcal{H}} |L_{\mathcal{D}}(\mathbf{h}) - L_S(\mathbf{h})|\right] \geq \frac{1}{5}$$

Therefore $(\mathcal{H}, \mathcal{Z}, \ell)$ *does not have the uniform convergence property.* □

For the learning problem presented here, we were able to show that there exists some hypothesis $\mathbf{h} \in \mathcal{H}$ for which the true risk does not converge to the empirical risk as $m \to \infty$. We can sharpen this example by exhibiting a problem for which the empirical risk minimizer $\mathbf{h}^*$ also exhibits this problem.

Consider the problem $(\mathcal{H}, \mathcal{Z}, \ell')$ where $\mathcal{H}$ and $\mathcal{Z}$ are as in Theorem 4, but the loss function is now

$$\ell'(\mathbf{h}, (\mathbf{x}, \boldsymbol{\alpha})) = \left\|\sqrt{\boldsymbol{\alpha}} * (\mathbf{x} - \mathbf{h})\right\|^2 + \eta \sum_{j=1}^{\infty} b_j (\mathbf{h}_j - 1)^2 \quad (4)$$

where $\eta = 0.01$ and $\{b_j : j \in \mathbb{N}\}$ is any set of positive numbers such that $\sum b_j = 1$.

The new loss function ℓ' is $(2 + 2\eta)$-smooth and since the additional term is strictly convex, ℓ' is strictly convex. Therefore, for any training sample S of any size,

$$\mathbf{h}^* \in \arg\min_{\mathbf{h} \in \mathcal{H}} L'_S(\mathbf{h})$$

is unique, where $L'_S(\mathbf{h})$ is the empirical risk of the hypothesis $\mathbf{h}$ with respect to the new loss function ℓ'. Assume that $\mathcal{D}$ is a product measure on $\mathcal{B}_{1/4} \times [0,1]$ satisfying conditions (1) and (2). In particular, $\mathbf{x}$ and $\boldsymbol{\alpha}$ are independent random variables.

Lemma 3. *Let* S *be a i.i.d. sample of size* m *of* $\mathcal{Z}$ *according to* $\mathcal{D}^m$. *If* $\mathbf{h}^* \in \mathcal{H}$ *is the unique empirical risk minimizer of* ℓ', *then a.s.* $\|\mathbf{h}^*\| = 1$.

Proof 3. *First consider the unconstrained optimization problem of finding*

$$\mathbf{h}^*_{UC} \in \arg\min_{\mathbf{h}} L'_S(\mathbf{h})$$

For any training sample S *of size* m, *a.s. there exists a coordinate* j *such that* $\boldsymbol{\alpha}_j^{(i)} = 0$ *for all* $i \in [m]$. *Thus only the second term in* (4) *depends on* $\mathbf{h}_j$. *Since* $\mathbf{h}^*_{UC}$ *is the unique minimizer of* L'_S, *we obtain* $\mathbf{h}^*_{UC,j} = 1$. *Consequently,* $\|\mathbf{h}^*_{UC}\| \geq 1$. *It follows that in the constrained case where* $\mathbf{h} \in \mathcal{H}$, *we must have* $\|\mathbf{h}^*\| = 1$. □

We will denote by $L'_{\mathcal{D}}(\mathbf{h})$ the true risk of hypothesis $\mathbf{h}$ under the loss function (4).

Theorem 6. *Let* S *be a i.i.d. sample of size* m *of* $\mathcal{Z}$ *according to* $\mathcal{D}^m$. *Let* $\mathbf{h}^* \in \mathcal{H}$ *be the unique empirical risk minimizer of* ℓ', *and let* $L^* = \min_{\mathbf{h} \in \mathcal{H}} L'_{\mathcal{D}}(\mathbf{h})$. *Then a.s.*

$$|L'_{\mathcal{D}}(\mathbf{h}^*) - L^*| \geq \frac{1}{5}$$

Proof 6. *By Lemma 3,* $\|\mathbf{h}^*\| = 1$. *We write*

$$\begin{aligned} L'_{\mathcal{D}}(\mathbf{h}^*) &= \mathop{\mathbb{E}}_{(\mathbf{x}, \boldsymbol{\alpha})}[\ell'(\mathbf{h}^*, (\mathbf{x}, \boldsymbol{\alpha}))] \\ &\geq \mathop{\mathbb{E}}_{(\mathbf{x}, \boldsymbol{\alpha})}\left[\sum_{k=1}^{\infty} \boldsymbol{\alpha}_k (\mathbf{x}_k - \mathbf{h}^*_k)^2\right] \end{aligned}$$

since $\mathbf{x}$ *and* $\boldsymbol{\alpha}$ *are independent with our choice of* $\mathcal{D}$:

$$\begin{aligned} &= \sum_{k=1}^{\infty} \mathop{\mathbb{E}}_{\boldsymbol{\alpha}}[\boldsymbol{\alpha}_k] \mathop{\mathbb{E}}_{\mathbf{x}}(\mathbf{x}_k - \mathbf{h}^*_k)^2 \\ &= \frac{1}{2} \mathop{\mathbb{E}}_{\mathbf{x}}[\|\mathbf{x} - \mathbf{h}^*\|^2] \\ &\geq \frac{9}{32} \end{aligned}$$

On the other hand,

$$\begin{aligned} L^* &= \min_{\mathbf{h} \in \mathcal{H}} L'_{\mathcal{D}}(\mathbf{h}) \\ &\leq L'_{\mathcal{D}}(\mathbf{0}) \\ &= \mathop{\mathbb{E}}_{(\mathbf{x}, \boldsymbol{\alpha})}[\ell'(\mathbf{0}, (\mathbf{x}, \boldsymbol{\alpha}))] \\ &= \mathop{\mathbb{E}}_{(\mathbf{x}, \boldsymbol{\alpha})}[\left\|\sqrt{\boldsymbol{\alpha}} * \mathbf{x}\right\|^2] + \eta \\ &\leq \mathop{\mathbb{E}}_{(\mathbf{x}, \boldsymbol{\alpha})}[\|\mathbf{x}\|^2] + \eta \\ &\leq \frac{1}{16} + \eta \end{aligned}$$

Since $\eta = .01$, $|L'_{\mathcal{D}}(\mathbf{h}^*) - L^*| \geq \frac{9}{32} - \frac{1}{16} - .01 > \frac{1}{5}$. □

Theorem 6 says that for any product measure $\mathcal{D} = \mu \times U_{[0,1]}$ on $\mathcal{B}_{1/4} \times [0,1]$ where $U_{[0,1]}$ is the uniform distribution on $[0,1]$, then a.s. the unique empirical risk minimizer $\mathbf{h}^*$ of $(\mathcal{H}, \mathcal{Z}, \ell')$ performs much worse than the population optimum L^* and therefore does not converge to it as $m \to \infty$. Thus this problem is not learnable under the ERM rule, although we already showed that is learnable under a RLM rule.

V. Conclusion

Theorems 5 and 6 are generalizations of Example 4.1 in [5]. In that paper, the authors only presented a single distribution $\mathcal{D}$ concentrated on $\mathbf{0} \in \mathcal{B}$ where the uniform convergence property fails. We have presented a very rich family of distributions over $\mathcal{B} \times [0, 1]$ where the gap between the empirical risk and the true risk is bounded away from 0 a.s. for any sample size m. The only restrictions on $\mathcal{D}$ that we have imposed are a concentration of $\mathcal{D}$ in a smaller ball of radius $1/4$, and a corresponding distribution of independent, uniform Bernouilli random variables on the $\boldsymbol{\alpha}$ variable. We can even relax some conditions on the problem we have studied here. For example, we can ask that the hypothesis class $\mathcal{H}$ is a bounded convex set only. This family of distributions for the weighted center of mass problem also shows that "pathogenic" distributions on $\mathcal{Z}$ where the uniform convergence property fails are much more common than originally thought.

References

[1] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth, "Learnability and the Vapnik-Chervonenkis dimension," *J. Assoc. Comput. Mach.*, vol. 36, no. 4, pp. 929–965, 1989.

[2] V. N. Vapnik and A. Y. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," in *Measures of complexity*, pp. 11–30, Springer, Cham, 2015. Reprint of Theor. Probability Appl. **1**6 (1971), 264–280.

[3] M. J. Kearns, R. E. Schapire, and L. M. Sellie, "Toward efficient agnostic learning," *Machine Learning*, vol. 17, pp. 115–141, 1994.

[4] B. K. Natarajan, "On learning sets and functions," *Machine Learning*, vol. 4, pp. 67–97, 2004.

[5] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan, "Learnability, stability and uniform convergence," *J. Mach. Learn. Res.*, vol. 11, pp. 2635–2670, 2010.

[6] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning - From Theory to Algorithms.* Cambridge University Press, 2014.

[7] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler, "Scale-sensitive dimensions, uniform convergence, and learnability," *J. ACM*, vol. 44, no. 4, pp. 615–631, 1997.

Dr. César Garza is an assistant professor of Mathematics & Statistics at the University of Houston-Downtown. His research focus is on the area of convex learning problems. He has supervised student research projects in PAC learnability. He has also published papers on the construction of hyperkähler metrics through Riemann-Hilbert problems and he is the author of a monograph in the area of low-dimensional topology about hyperbolic knots with toroidal surgeries.

Effects of Changes in Accounting Standards On Financial Disclosure for MFIs: The Case of MFIs in Cameroon

1. Dr Fidelis Akanga

Bournemouth University

Department of Accounting, Finance and Economics

Fern Barrow

BH12 5BB

Poole, UK

2. Dr Jacob Agyemang

University of Essex

Wivenhoe Park

CO4 3SQ,

Colchester, UK

jagyem@essex.ac.uk

3. Dr Widin B. Shav'en

University of Leicester,

London Road

LE2 1RQ

Leicester, UK

wbsv1@leicester.ac.uk

4. Dr Dominic Roberts

MacEwan University,

104 Ave NW, Edmonton, AB T5J 4S2,

Canada

robertsd43@macewan.ca

Abstract

The aim of this paper is to evaluate if changes in accounting standards for MFIs do improve financial disclosure, operational and financial performance

This paper make three important contributions; Firstly, it contributes to accounting literature by looking at how changes in accounting standards in an emerging economy can affect MFIs financial disclosure and performance. Secondly, it contributes to microfinance literature by explaining how changes in accounting standards within the MFI sector affects financial disclosure and activities of MFIs. Thirdly, the explains how changes in accounting standards can affect MFIs meeting their social needs and poverty reduction.

This paper is based on 35 indepth interviews with Managers and Accountants of MFI consortia in Cameroon.

Our findings show that changing accunting standards for MFIs without any well developed financial markets do not benefit the MFIs and do not improve financial disclosure. Secondly, accounting change is not of prime importance to MFIs given their sizes, their stakeholders and the nature of their activities. Rather than changing accounting standards for MFIs, MFIs should be provided guidelines on how their financial statements be presented to help imrpve their performance.

1. Introduction

In 2011, Cameroon and 14 Francophone states, one Spanish speaking state (Equatorial Guinea), and one Portuguese speaking (Guinea Bissau) state in Africa opted for a new accounting treaty, the Harmonisation of Business Laws in Africa [French acronym, OHADA] for enterprises in Africa. This change according to [36], is “to harmonize and modernize business laws in Africa so as to facilitate commercial activity, attract foreign investment and secure economic integration in Africa” as stipulated “within the framework of the objectives of the New Partnership for Africa’s Development (NEPAD) agreed in June 2002 at the G8 Kananaskis summit”. It is worth noting here that, this change was for all African states. But Anglophone states have not signed up to this treaty due to linguistic problems, Article 42. Article 42 states that the working language of the treaty is French [36] even though there is considerable progress at translating this document into English.

The Banking Commission for Central African States [french acronym, COBAC] was drafted in 2002 to regulate the Microfinance sector in the Monetary Union of Central African States [French acronym, CEMAC]. COBAC was drafted in following a period of huge losses by Microfinance Institutions (MFIs) in the CEMAC region. In 2002, COBAC instituted regulations on Cooperative Societies (COBAC, 2002) and in 2011, COBAC made mandatory the OHADA Treaty for MFIs [2]. The aim of COBAC was to institute good governance practices, increase accountability, strengthen risk mechanisms and improve risk strategies, and improve financial disclosure across the sector. However, the outcomes have been mixed. In terms of improving governance practices, the changes instituted shareholders within the MFIs resulting to hybrid-institutions where managers of MFIs became more focusd on meeting the demands of shareholders at the expense of the poor [39]. In terms of accountability, the COBAC regulations and due to lack of proper monitoring and supervision, MFIs opened multipe lines of accountability that allow them meet the regulatory requirments [3]. In the area of risk, the COBAC regulations allowed managers of MFIs to price risk differently to help them meet the demands of the shareholders [4] & [12]. For over a decade since the COBAC regulations and 9 years since OHADA treaty were made mandatory for MFIs in Cameroon and CEMAC, little is known if at all the changes in accounting standards have improved financial disclosure within the microfinance sector in Cameroon especially as MFIs in this region have adopted for OHADA instead of International Financial Reporting Standards (IFRS) [72].

The reason for the changes in accounting standards for MFIs in Cameroon is to help them communicate their activities, financial performance and position to potential investors [71]. IFRS is therefore seen as one of the best methods to help these MFIs not only do this but also increase transparency [42]. Unfortunately, there is no uniformity when it comes to adopting these accounting standards. To [79], MFIs should be allowed to adopt different standards and IFRS if they operate in an area where the accounting standards are not well-developed or if there are no national accounting standard. [72] argue that this has led to a situation where the adoption of these standards varies from country to country and in some cases, others adopt regional standards such as the case with OHADA for MFIs in Cameroon. As a result, MFIs are now using different standards to prepare financial reports for their different stakeholders [74], with some using national accounting standards when reporting to regulatory authorities while others use IFRS when reporting to external donors and international investors [5].

According to [73], because of increase commercialisation of MFIs and some taking the for-profit status, adopting IFRS provides incentive for investors wishing to diversify their portfolio to invest in MFIs [15]. [40] and [32] argue that adopting new accounting standards such as IFRS and preparing financial statements following these standards lowers information risk and capital cost. [15] argue that, adopting new accounting standards leads to improved quality of financial reports and accounting figures.

The aim of this paper therefore is to evaluate the effect of changes in accounting standards for MFIs and if these changes have improved financial disclosure. Previous studies have shed light on the reasons why public and privately listed companies in developed countries will adopted new accounting standards [35], [6], [28], [41], [13], [32], [5]. With the execption of studies by [1] and [73] who looked at the reasons why MFIs in emerging economies choose to adopt new accounting standards, there are no studies that look at the effect of adopting these standards by MFIs operating in emerging economies. The paper addresses the limitation by looking at the effect of adopting these accounting standards in the context of a developing economy. Following, this paper intends to answer the question; Does adoptng new accounting standards improve financial disclosure for MFIs in an emerging economy? This paper also makes the following important contributions; Firstly, it contributes to accounting literature by looking at the outcome of changes in accounting standards on MFIs financial disclosure in an emerging economy. Secondly, it contributes to microfinance literature by looking at how changes in accounting standards can affect financial disclosure and activities of MFIs. Thirdly, the findings of the paper brings out relevant poverty mitigating policy implications as in part highlight how changes in accounting standards can affect MFIs in meeting one of the metrics of the UN Sustainable Development Goals (SDGs) of ending poverty by 2030 [45].

This paper is structured as follows; The second section looks at the history of MFIs in Cameroon; The third section, explains the accounting changes in Cameroon and their effects on organisations. The fourth section presents the research methods; The fifth section data analysis of findings, and the sixth concludes by highlighting by the policy related and other contributions and limitation of the paper.

2. Brief history of Microfinance in Cameroon

MFIs emerged to fill the gap created by mainstream financial institutions as an alternative means of providing capital to the low-income people such as "loans without collateral, group lending, progressive loan structure " [78] to help them invest in income generating activities thereby contributing to eradicating poverty [45]. MFIs have since extended their activities to cover a broader range of activities such as savings, insurance, money transfer services; in general, and development lending [53].

Microfinance as it is in Cameroon today can be traced back to 1963 where it was introduced in Njinikom (a suburb in the North West Region) by Rev. Father Anthony Jansen, a Roman Catholic priest originally from Holland [56]. The Rev. Father discovered that the local communities were not only selling their cash crops before the harvesting season but were also involved in localised thrift and loan societies popularly referred to as "njangis" charging very high interest rates for loans used to buy fertilizers and pesticides. He then brought together sixteen members of his congregation who started with a discussion group referred to as "St. Anthony Discussion Group". These sixteen members started with a savings scheme which has since been transformed into (MFIs) operating today across Cameroon [3].

The microfinance sector in Cameroon really gained grounds after the 1980s following the banking crisis in Cameroon because of major failings linked to the government interventions in the banking sector, inadequate management of the banking sector and lack of any credible enforcement regime of the banking regulations [2]. MFIs took advantage of that gap in the market especially as Cameroonians were in need of financial services such as savings schemes that were no longer available and established itself as a major option for the poor who in most part were segregated by the mainstream financial institutions [69].

3. Related literature on MFIs and Accounting Standards

Accounting play different roles within organisations. [49] argues that "accounting has provided an operational and influential language of economic motives, its calculations had infused and influenced important policy decisions, and the visibilities it created played an important role in making real particular segmentations of the organisational arena. Accounting not only reflected the organisation as it had been but also played a not insignificant role in positively making the organisation as it now is". To [58], accounting has been used to create a pattern that are tied to modes of legitimation and relations of power within organisations.

Accounting is seen as providing a constructive method in shaping the role played by actors within organisations by “spreading concepts like value for money, accountability, efficiency, effectiveness, turning them into new shared meanings and values” [20]. Accounting plays a vital role in organisations by altering the organisational and social life therefore influencing perceptions of actors, changing the language of the organisational actors and infusing dialogue, and thereby changing the ways in which “priorities, concerns and worries, and new possibilities for action are expressed” [50]. Accounting also plays a significant role in the creation of organisational domain [49] and governable person [66]. Unfortunately, “little is known about how accounting systems are created and developed” [76]. This has led to others believing that accounting is a political, social, and economic process fabricated by both internal and external actors to change organisations [50], & [52]. [21] argue that, accounting has become a tool used by “powerful actors pursuing their own interests – political and economic”. This has led to changes in accounting standards highly contested especially as the reasons for the change is likely to be linked to various organisational pressures and rationales [79].

Changes in accounting standards within organisations are not new as it is often the result of some external forces acting on the organisation and internally by management to progress own agenda [52]. Factors such as social, political, and economic are today widely seen as significantly contributing to changes in accounting standards within organisations [50]. According to [83], these changes have often been on how organisations should organise their accounts, requirements to buy new software, changing their accounting rules and procedures such as financial disclosure and formal responsibilities and giving instructions to collect new data. The reason behind these changes in accounting standards within organisations are to align accounting with organisational, and economic contexts [50].

However, the role of accounting "in the emergence of organisations as we now know, the external and internal boundaries which they are conceived of having .... have been subjected to little investigation. Relatively little consideration has been given to the ways in which accounting changes has become implicated in, and, in turn, shape" [51] financial disclosure within organisations and Microfinance Institutions (MFIs) in particular.

“Accounting reforms have been rare in Cameroon … with much of its pressure for reforms is coming from external forces such as, World Bank and other donor agencies” [68]. However, viewing the important role accounting plays especially with economic development, [67] argues that Cameroon had to adopt accounting reforms and integrate International Accounting Standards (IAS) in order to attract foreign investments.

These changes in the financial sector in Cameroon has not been followed by swift reforms in accounting standards in Cameroon as to bolster the microfinance sector. The reason for the slow reforms in accounting standards in Cameroon is widely attributed to the fact the Cameroon’s accounting and financial standards are aligned more to CEMAC accounting plan which is derived from the French “Plan Comptable General” with a similarity of 64% to France’s system [67], with its “pre-planned accounting codes” [68]. As a result, [42] argue that accounting reforms in Cameroon in most cases is accidental.

Over the past 60 years of the political history of Cameroon, accounting reforms have been slow and accidental [42] & [68]. Firstly, up to 1961, Chapter 37 of the Laws of the Federation of Nigeria and Lagos 1958 was the accounting standard used by the then British Cameroons. This accounting plan according to [61] was modelled on the British Companies Ordinance of 1922. They argue that the specificity with this accounting plan did not require organisations operating in British Cameroons to keep accounting and bookkeeping records. On contrary, the French Cameroon was using the Ordonnance de Commerce of 1673 put through by Jean-Baptiste Colbert during the reign of Louis XIV, and the Napoleonic Commercial Code of 1807 [34]. Ordonnance de Commerce according to [34] had an elaborate codification system and organisations operating in French Cameroon were required by law to keep accounting records.

In 1965 and following unification of the two Cameroons (British Southern Cameroons and French Cameroons in 1961), the two accounting plans were changed to Plan Comptable General [33] & [73]. Then came The African and Malagasy Common Organization (French acronym OCAM) in 1970. OCAM’s Accounting Plan was derived from the French Plan Comptable of 1947 and 1957. The difference between the two was at the level of the codification of the principal accounts, the introduction of financial statements such as the fund accounts, an income statement section to measure gross profit and value added, consolidation accounts, prescribed measures for asset valuation and income measurement, and the requirement for notes explaining the accounts and the accounting policies used in preparing the statements [50]. OCAM Accounting plan had an elaborate codification system (in the form of index card) that specified which code had to be credited or debited [50]. According to [50], it took accounting to different levels by emphasising on both micro and macro aspects of accounting. [33] argued that it "was very much in tune with the spirit of the British Corporate Report (ASC, 1975), the Canadian Stamp Report (CICA, 1980) and the Trueblood Report (AICPA, 1973)".

However, in 2011, OCAM was replaced by OHADA. According to [35], OHADA is a blend of the Anglo-Saxon model of accounting with the French accounting system approach by codifying some of the provisions of International Financial Reporting Standards (IFRS) and incorporating them as articles within the framework of OHADA in line with the French civil law tradition "wherein codes and statutes are highly structured and systematized" [35]. [68] argues that adoption of OHADA by Cameroon came after long hesitation by Cameroon in adopting the Anglo-Saxon accounting approach because of the versatility it carries.

Since 2011, the OHADA accounting treaty has been mandatory for all MFIs operating in the CEMAC region [2]. This treaty is divided into four major parts (General guidance on formation of Cooperative Societies in Africa, guidance as to what should be done in case of dissolution of any Cooperative Society, penalties and sanctions, and other issues relating to the activities of Cooperative Societies in Africa), and into 390 articles [69]. Prior to this coming into effect, in the CEMAC region, there were similar regulations put in place by the Banking Commission for Central African States (COBAC) regulations on Cooperative Societies in 2002 [23].

Accounting change is often seen as being exogenously driven and a result of regulations and some cosmetic behaviour of external forces to constrain and shape organisations [37]. The reason behind these changes is to provide an easy method of transparent and comparability of cross-country firm's financial reporting and disclosure [50]. Unfortunately, these changes in accounting standards have been seen as being burdensome for managers as these do not allow them to communicate effectively with investors or permit them show the "benefits of investments in quality improvements, human resource development programmes, research and development and customer service" [50].

[46] argue that, the effect of any changes in accounting standards to improve the quality of financial reporting does not yield any benefits compared "to the effects offerees such as managers' incentives, auditor quality and incentives, regulation, enforcement, ownership structure, and other institutional features of the economy in determining the outcome of the financial reporting process". [24] and [59] argue that changes in accounting standards alone do not translate to improved quality of financial disclosure. [10] argue that, as a result to improve on financial disclosure through changing accounting standards, this had led to undermining key institutional features such as manager and auditor incentives therefore not guaranteeing financial disclosure. [9] argue that changes in accounting standards are promoting "rules-based" approach to financial reporting over a "principles-based" approach. This has resulted to a situation where organisations place more emphasizes on accounting disclosures that fairly represents the company's financial position, as opposed to the former, which emphasizes on compliance over substance [46]. This has resulted in accounting moving from a customised and "inward-looking" philosophy through to a "control-based" approach to today's "market-based" approach [51].

4. Research Methodology and Theoretical Framework

The main research method for this research is case study and qualitative data was collected through in-depth interviews, observations, and document review (a combination of both inductive and deductive approaches). Through case study, we were able to investigate how changes in accounting practices leaves blurred boundaries within the MFI sector, their activities, and the poor [85]. Changes in accounting practices has been a medium and an outcome of articulating contradiction [78] and alternative approaches [25]. Therefore, through case study, and using the case of MFIs in Cameroon, this provided us with the opportunity to investigate how changes in accounting standards can affect financial disclosure [62] & [77]. Through interviews, we were able to gain an understanding of the changes in accounting standards as perceived by the actors, how these changes have affected ideas within the organisations, meanings, interactions and implementation of these changes. Through document review, we looked through official accounting documents such as budgets, financial reports and most accounting documents which represents these accounting change. We then analysed these documents to ascertain any changes that might have occurred overtime which helped us understand the evolution of these changes, the source of this change and the relationship between both the internal and external actors. We also observed the different communications concerning these changes. Our aim was to understand how the different actors adopted these new changes, their perception and how these changes affected decision making by actors. We were also able to go through meeting notes of managers during meetings with shareholders and other key members of staff. This provided us the opportunity to understand the reaction of the shareholders, members of staff and managers perceptions to these changes and possible disruptions to financial disclosure [7] & [43].

This research is based on 35 in-depth interviews (semi-structured) with managers (20) and accountants (15), observations and document review. In-depth interviews were conducted in Cameroon between June 2019 and March 2020 with managers and accountants of MFIs of the two largest MFI consortiums in Cameroon, Cameroon Cooperative Credit Union League (CamCCUL) and Mutuelle Communautaire de Croissance (MC^2). The interviews were based on the problems the interviewees encounter with the changes in accounting practices in Cameroon, problems with monitoring and implementation of these changes, the choice of financial disclosure used within their organisations, and the effects on their activities and the society as a result of these changes in accounting practices and financial disclosure.

The thirty-five interviewees were purposively selected using the key informant technique from 45 that accepted to participate in the research [60] and the criteria highlighted below. The choice of 35 respondents was based on how long the respondent has been working within the microfinance industry in Cameroon (a minimum of five years), the respondent's knowledge of the subject matter, their influence within the area and, the participant's availability. The same questions were asked to all respondents and the respondents were allowed to express themselves freely. The researcher's role was moderating the interview process and where necessary, and followed up with a probing question to clarify on a subject discussed but not explicit by the respondent. The interviews ranged from thirty minutes to one hour.

The data was transcribed and analysed using thematic analysis. Thematic analysis is used in qualitative research or data collection to analyse different classifications and themes related to the data and thus provided a means to illustrate in greater detail on diverse subjects by using different forms of interpretations [17]. In our case thematic analysis is considered appropriate for two major reasons; Firstly, it allowed us the opportunity to relate the concepts, behaviours, actions and thoughts of our interviewee and the data [55]. Secondly, since our data was collected using interviews, observations and document review, thematic analysis allowed us to adopt the flexibility of either using inductive or deductive approach [64]. Through the inductive approach, which most of the data was collected, we started with a precise exploration of our topic/question(s) and then moved to generalisations and then theory. Thematic analysis allowed us to link our data to the different themes generated [70]. Through deductive approach, thematic analysis allowed us to construct meaning from the different themes emanating from participants' opinions or feedback during the study.

In our case, data was gathered using different methods with different participants at different locations. Thematic analysis according to [18] provides researcher with the means to identify, analyse, organise, describe, and report the different themes found within a data set. [17] argue that, in such cases, thematic analysis provides the most effective method to produce and present data that will reflect the reality. Thematic analysis provides us the means to investigate how the current accounting practices have been influenced by actors and then highlight the apparent differences or similarities in the outcome of the financial disclosure by organisations.

Accounting change is influenced by a "multiplicity of agents, agencies, institutions and processes" [65]. Change within organisations can originate from either internal or external actors and is shaped by the internal or external forces [20]. As a result, the role of accounting in shaping organisational change has been studied using different theoretical institutional perspectives. New Institutional theorists have argued that changes in accounting practices within organisations are often driven by powerful organisations [63], [75], and [80] and the result of new regulations or behaviours of external powerful agents [83]; [37], [57]. They argue that because of institutionalisation and diffusion of power, this affects the decisions made by those within the organisations and thus leads to a situation where this converges into isomorphisms. On the other hand, old institutional theorists believe that changes in accounting within organisations is the result of accounting rules and routines that have become stabilised and transformed into new ones through the process of institutionalisation and deinstitutionalisation [19]. These two perspectives have failed to explain how the pace and outcome of this accounting change interacts with the internal dynamics of the organisation facing similar pressures [84]; [30]. [30] argue that, these two perspectives do not provide explanation as to how these accounting practices have become institutionalised, changed, or deinstitutionalised within organisations. To [22], the two institutional perspectives are more focused on the organisational field and failed to explain the role played by powerful agents, those with special interest on the organisation and the role played by the political nature of the organisation.

In our case, for us to evaluate the effect of changes in accounting standards on MFIs and if these changes have improved financial disclosure, we will use the archetypes theory [44],[45]. According to [56] change represents "the shift between different archetypes and comprises both structures and systems and beliefs and values". To [20], change can be formal or informal and can result to different outcomes "(radical vs. incremental), paces (i.e., evolutionary versus revolutionary) and levels (structures and systems versus ideas and values) of change, adopting a micro-level perspective (the organisation)" [56].

Given the fact that there are different external agents responsible for accounting change within organisations, organisations adopt different methods to respond to these pressures [27]. Organisations therefore provide that domain for these different agents to interact in pursuing their goals and interest by either willingly or unwillingly distorting the processes that existed within the organisation and turning the organisation into their personal weapon [30]. By using the Archetypes theory, we are creating an internal process for us to interpret the external environmental pressures resulting from the external and internal agents [45]. To [56], the Archetype theory provides the possibility for organisations to respond differently to any change stimuli. [56] further argue that, the Archetype "theory provides a finer-grained definition of the final outcome of change, in terms of shift in both structures and systems and related ideas and values". By using Archetype theory, we differentiate between the systems and structures that are undergoing change and the different values and beliefs used by actors to define if change has been achieved [53].

5. Data Analysis and Presentation of Findings

We started by asking the managers to understand the role of the supervisory and monitoring authorities in Cameroon on their choice of accounting standards. According to Theodore, "the role of the supervisory and monitoring authorities in Cameroon is to make sure that whatever accounting standard we use should help MFIs produce accurate, relevant and timely financial reports". Jean-Jacque stated that "the role of the supervisory authorities is to ensure that we adhere to OHADA accounting standards in preparing our financial statements". To Marie, "their role is also to ensure that our financial statements submitted to the Ministry of Finance on a yearly basis represent the true value of our institutions". The reason why financial statements are submitted yearly to the Ministry of Finance is to avoid managers falsifying their accounts and financial misrepresentation [38]. [38] argue that it is important that supervisory authorities compare these financial statements with those of previous years to see if the managers have restated their financial position as this is direct evidence of accounting manipulation. Joana argued that the reason why these accounts are submitted to the Ministry is to make sure that "the accounts conform with the OHADA treaty". [79] argue that supervisory authorities need these financial statements to enforce compliance by organisations. Kingsley argued that "the reports are often required by shareholders during board meetings". These reports according to William "are used to determine our performances and pay progression". This is consistent with the argument by [86] that owners of SME are increasingly dependent on financial statements to make important decisions about their organisations. However, [73] argue that, there is an increase discrepancy on how financial statements are used by different stakeholders. These discrepancies in the use of financial statements by different stakeholders according to Emilia "is because the regulations in place have brought different stakeholders and especially shareholders who are more profits focused". To John, "the use of financial statements by stakeholders varies with level of education". [86] argue that owners who are knowledgeable about financial statement are better placed on making good decisions about the profitability, liquidity and the different risks their organisations are exposed to compared to those with little or no idea of financial statements.

Our next question was to find out given that users of accounting information and their needs is broad, if they require different accounting information. Almost all the accountants of the MFIs under the two MFI (CamCCUL and MC^2) consortia in Cameroon, agreed that their users are "branch managers, MFI managers, board members, external auditors, regulatory and supervisory authorities, and investors" and Sebastian stated that they "have different information needs". Thomas informed us that "knowing the users and their information needs allow them build their systems and reports to facilitate with providing information when needed". According to [29], the role of accountants within MFIs is very important in three ways; designing the accounting system used, are the primary users of this system to make sure that the right information is provided to the user in the right format, time and size, and play the role of internal auditor. For this reasons, [29] argue that the communication method used by the accountant to communicate information is critical for the success of MFI.

Our next question was to know the "systems" they are using to meet their needs. According to Dickson, "due to MFIs being of different sizes, they have invested in different software". It should be noted here that the accountant was not willing to provide further details of the software. To Mercy, "the software they use is provided by Afriland First Bank". [81] argue that, the reason for this software is to help these institutions organise their accounts in order to provide accurate and timely information to their stakeholders [40].

Our next question was to get the view of our respondents on the use OHADA accounting treaty compared to IFRS to produce financial statements. According to Mercy, "OHADA is a regional standard based on the accounting of the different countries in the CEMAC zone and we are required to prepare our financial statements following the treaty". On the other hand, Dickson argued that "OHADA is more of the old French Plan comptable General. This means that financial statements have some differences between OHADA and IFRS due to codification of accounts under OHADA". However, [31] argue that, the use of domestic accounting standards in producing financial reports reduces the quality of such reports. To Dickson, "there are differences between OHADA accounting treaty and IFRS and does not allow much room for comparability". The problem with issues of comparability according to [8] is that OHADA treaty was developed without any prior development of regulations guiding the CEMAC market. This according to Elizabeth, "is further compounded with differences in interpretation due linguistic differences between the Anglo-Saxon and the French countries" (Cameroon not excluded as the country uses both languages). As a result of these issues, [31] argue that without any changes to the market in an area, adopting new accounting standards will not yield any significant benefits.

Our next question was to find out from our respondents if they use any other financial standard other than OHADA to produce their financial statements. Theodore, John, Joana and Lucas agree to using IFRS from time to time. On their part, Sebatian, Thelma and Ebogo stated that they use only OHADA. The reason for these differences according to Benjamin is "to lack of any version of OHADA treaty in English". According to [36], because of no authoritative version of the OHADA Treaty in English is the reason why many countries in Africa with the Anglo-Saxon culture have not adopted this treaty. Thomas argued that they "use IFRS at times to avoid understating the value of our MFIs". On his part, Lucas argued that "we are category one MFIs and have shareholders who want to see a return on their investment. For this reason, we have to produce our accounts to allow investors make informed decisions". The essence of a unified accounting standard is to allow firms to produce financial reports that best reflect the firm's economic position and performance [11]. [11] & [43] argue that, national accounting standards allow managers of MFIs "to manage earnings thereby decreasing their accounting quality". [70] argue that firms using national accounting standards are likely to report lower net income, less cashflows and lower equity value for share prices. [16] argue that, MFIs need to follow international accounting standards in preparing their financial statements if they require more capital which is controlled by those outside of the organisation. However, [78] argue that MFIs do have this special characteristic of using financial business model to support social mission of providing soft loans to the poor. To [48] & [78] their economic or financial performance should not be judged based on the level of financial disclosure, but if they meet the two minimum requirements; "Is the information essential for understanding the core condition of an MFI and its potential to move beyond reliance on scarce subsidized funding? In present practice, is the information frequently missing from MFI financial statements?"

Given the differences in the use of accounting standards, we asked Elizabeth, Sampson, Joana and Lucas about enforcement. Sampson argued that "monitoring and enforcement of the use of accounting standard is with the Ministry of Finance". To Elizabeth, "the Ministry does not have enough staff to strictly enforce the accounting standards used by MFIs". Due to laxed enforcement of accounting standards in Cameroon, [11] argue that, MFIs are now adopting "principle - based approach". [81] argue that, government agencies lack the capacity and are unwilling to regulate MFIs because MFIs are of different types, legal structure and use different methodologies to meet their social objective [13], [3]. To [16], without any strict monitoring and enforcement on the nature of accounting standards used, MFIs are unlikely to grow and achieve that degree of financial self-sufficiency, outreach and impact because of those outside being unable to properly asses the institutions because of lack of audited financial statements. However, [78] argue that, "even if financial statements are not audited, … disclosure guidelines are intended to apply, especially when financial statements are used to present an MFI's financial condition to outsiders such as donors or investors". According to Emilia, "despite the Ministry not having the manpower to supervise and monitor our activities, we do submit yearly financial statements to stay within the law". According to John, "being part of the CamCCUL Network, we are required by our umbrella organisation, CamCCUL to submit our financial statements to the Ministry of Finance". [82] argue that organisations that voluntarily disclose their financial reports means that such information are truthful, adequate, and represent a transparent process where their shareholders and investors can easily access this information. [14] & [54} argue that, such voluntary disclosure by MFIs increases public trust on these institutions, improves on the quality of decisions made by shareholders and investors, and increases donations. Unfortunately, going through documents of eight MFIs, we observed, because of these voluntary disclosures, they ordered to pay corporation taxes because of their profit margins.

Our next question was to find out the effect of the changes in accounting standards on the activities of MFIs. To Isaac, "the greatest problem with the requirements of OHADA is for us to produce these accounts. Unfortunately, most of those who need these financial statements to make decisions are not familiar with these financial statements let alone appreciate or understand these financial statements". On his part, Ashu, argued that "without understanding these financial statements, it is a challenge for the shareholders and supervisory authority to use these accounts to monitor and manage MFIs". [26] argue that, without the supervisory authority understanding the different financial statements, this often leads to confusion between the use of prudential and non-prudential regulations to regulate the activities of MFIs. Another problem with the changes in accounting standards according to Edmond is "the apparent lack of professionals who are able to prepare accounting statements complying with the OHADA Treaty". He further argued that this situation is due to "lack of training offered by the state prior to rolling out the treaty". To Roland, "the state should have started by getting every stakeholder involved in the policy making process and made sure everyone working within the financial sector, be they accounting associations, accounting and auditing professional bodies or donor organisations participated in the dissemination and adapting the OHADA treaty". On his part, Olivia argued that "without training on this treaty, training of staff on this treaty has been pushed to both the accountant and MFIs". She further argued that "this has led to MFIs incurring huge costs on the services of external auditors". To Oliver, "one of the problems with changes in accounting standards leaves MFIs with doubts on how to report certain assets and especially loans. Most big MFIs report their loans using accruals while smaller MFIs report their loans on a cash basis". As [47] argue, this shows the effect on changes on accounting standards cannot be uniform in both the Northern developed countries and Southern developing countries.

6. Discussion and Conclusions

Accounting change is important to help investors and others outside of the organisations to make informed decisions about the organisation. In the case of MFIs, accounting change is not of prime importance given their sizes, their stakeholders and the nature of their activities. What is important for MFIs are guidelines on how they can present their financial statements to meet the needs of their users.

In the case of MFIs in Cameroon, the OHADA accounting treaty was hastily drafted and implemented without any prior consideration on the effect on their activities. OHADA treaty has left institutions with more cost especially as there was not prior training and no authoritative version in English. Without well-developed capital market in the CEMAC region, there are little benefits in terms of improved financial disclosure from changes in accounting standards. Without proper training, users of the financial statements are unable to make useful decisions. It is therefore important that governments in the CEMAC area roll out training and not allow that in the hands of the MFIs.

Appendix 1: List of participants and their organisations

	Names	Gender	Organisation	Position
1	Theodore	Male	CamCCUL	Manager
2	John	Male	CamCCUL	Manager
3	Jean-Jacque	Male	MC^2	Manager
4	Sebastian	Male	MC^2	Accountant
5	Marie	Female	MC^2	Manager
6	Thomas	Male	CamCCUL	Accountant
7	Elizabeth	Female	CamCCUL	Manager
8	Ebogo	Male	MC^2	Accountant
9	Sampson	Male	CamCCUL	Manager
10	Joana	Female	CamCCUL	Accountant
11	Lucas	Male	CamCCUL	Accountant
12	Lucia	Female	MC^2	Manager
13	Mercy	Female	CamCCUL	Accountant
14	Marius	Male	CamCCUL	Manager
15	Manfred	Male	CamCCUL	Accountant
16	Dickson	Male	MC^2	Accountant
17	Thelma	Female	MC^2	Manager
18	Olivia	Female	CamCCUL	Accountant
19	Oliver	Male	CamCCUL	Accountant
20	Divine	Male	CamCCUL	Manager

21	Emilia	Female	MC^2	Manager
22	Ernest	Male	CamCCUL	Accountant
23	Benjamin	Male	CamCCUL	Manager
24	Isaac	Male	CamCCUL	Accountant
25	Roland	Male	MC^2	Accountant
26	Elsie	Female	CamCCUL	Manager
27	Mary	Female	CamCCUL	Accountant
28	Ashu	Male	CamCCUL	Accountant
29	Johnson	Male	CamCCUL	Manager
30	Edmond	Male	MC^2	Manager
31	Timothy	Male	CamCCUL	Manager
32	Kingsley	Male	CamCCUL	Manager
33	William	Male	CamCCUL	Manager
34	Lawrence	Male	MC^2	Manager
35	Miriam	Female	CamCCUL	Manager

References

1. Adah, A. (2014). "Inconsistency in the Adoption of IFRS by Nigerian Microfinance Banks". *Research Journal of Finance and Accounting, Vol. 5, Issue No. 14, pp. 161 – 167*

2. Akanga, F.K. (2016). "Governance of Microfinance Institutions (MFIs) in Cameroon; what lessons can we learn?" *Enterprise Development and Microfinance, Vol. 27, Issue No 3, pp. 219 – 235*

3. Akanga, F.K. (2017). "Microfinance accountability in Cameroon: a cure or a curse for poverty alleviation?" *Journal of Accounting and Organizational Change*, *Vol. 13, Issue No 1, pp. 112 – 130*

4. Akanga, F.K., Sha'ven, W.B., & Tauringana, V. (2020). "An Empirical investigation into the Risk Management Strategies of MFIs in Cameroon". *African Journal of Accounting, Auditing and Finance, Vol. 7, Issue No. 2, pp. 155 – 171*

5. André, P., Kalogirou, F., (2019). "IFRS adoption by UK unlisted firms: Subsidiary-versus group-level incentives". *Accounting Forum, Vol. 44, Issue No. 3, pp. 215 – 237*

6. Ashbaugh, H. (2001). “Non-US Firms’ accounting standard choices”. *Journal of Accounting and Public Policy, Vol. 20, Issue No. 2, pp. 129 – 153*

7. Ashraf, J. M. & Uddin, S. (2013). ”A Consulting Giant; a Disgruntled Client: A ‘Failed’Attempt to Change Management Controls in a Public Sector Organisation”. *Financial Accountability & Management, Vol. 29, Issue No. 2, pp. 186 – 205.*

8. Ball, R. (2001). “Infrastructure requirements for an economically efficient system of public financial reporting and disclosure”, *Brookings-Wharton papers on Financial services. Vol. 2001, pp. 121 – 169*

9. Ball, R. (2009). "Market and Political/Regulatory Perspectives on the Recent Accounting Scandals." *Journal of Accounting Research, Vol. 47, Issue No. 2, pp. 277 – 323*

10. Ball, R., Robin, A., & Wu, J.S. (2003). "Incentives Versus Standards: Properties of Accounting Income in Four East Asian Countries." *Journal of Accounting & Economics, Vol. 36, Issue No. 1-3, pp. 235 – 270*

11. Barth, M.E., Landsman, W.R., & Lang, M.H. (2008). “International Accounting Standards and Accounting Quality”. *Journal of Accounting Research, Vol. 46, Issue No. 3, pp. 467 – 498*

12. Bassem, B.S. (2012). “Social and Financial performance of Microfinance institutions: Is there a trade off? *Journal of Economics and International Finance, Vol. 4, Issue No. 4, pp. 92 – 100*

13. Bassemir, M., (2018). “Why do private firms adopt IFRS?” *Accounting and Business Research, Vol. 48, Issue No. 3, pp. 237 – 263.*

14. Blouin, M.C., Lee, R.L., & Erickson, G.S. (2018). “The impact of online financial disclosure and donations in non-profits”. *Journal of Non-profit and public sector marketing. Vol. 30, Issue No. 3, pp. 251 – 266*

15. Brière, M., Szafarz, A., (2015). “Does commercial microfinance belong to the financial sector? Lessons from the stock market”. *World Development, Vol. 67, pp. 110 – 125.*

16. Brown, P. (2011). “International Financial Reporting Standards; What are the benefits?” *Accounting and Business Research, Vol. 41, Issue No. 3, pp. 269 – 285*

17. Boyatzis, R.E. (1998). T*ransforming qualitative information: thematic analysis and code development*. Sage, Thousand Oaks, California.

18. Braun, V., & Clarke, V. (2006). “Using thematic analysis in psychology”. *Qualitative Research in Psychology, Vol. 3, Issue No. 2, pp. 77–101.*

19. Burns, J. & Scapens, R.W. (2000), “Conceptualizing management accounting change: an institutional framework”, *Management Accounting Research, Vol. 11, Issue No. 1, pp. 3 – 25*

20. Caccia, L. & Steccolini, I. (2006). “Accounting change in Italian local governments: What’s beyond managerial fashion?” *Critical Perspectives on Accounting, Vol. 17 Issue 2-3, pp. 154 – 174*

21. Carpenter, V.L., & Feroz, E.H. (1992). “GAAP as a symbol of legitimacy: New York State’s decision to adopt generally accepted accounting principles”. *Accounting Organisation and Society, Vol. 17, Issue No 7, pp. 613 – 43*

22. Carruthers, B.G. (1995), “Accounting, ambiguity, and the new institutionalism”, *Accounting,Organizations and Society, Vol. 20, Issue No. 4, pp. 313 – 328.*

23. COBAC (2002). Reglement CEMAC Relatif aux Conditions d'Exercice et de Controle de l'Activitite de la Microfinance. 01/02/CEMAC/UMAC/COBAC, Cameroon: COBAC

24. Coffee J.C, Jr. (2007). "Law and the Market? The Impact of Enforcement." *University of Pennsylvania Law Review, Vol. 156, Issue No. 2, pp. 229 – 311*

25. Chew, A. & Greer, S. (1997). “Contrasting world views on accounting: Accountability and Aboriginal culture”. *Accounting, Auditing & Accountability Journal, Vol. 10, Issue No. 3, pp. 276 – 298.*

26. Christen, R.P., Lyman, T.R., Rosenberg, R. (2003). *Microfinance Consensus Guidelines: Guiding Principles on Regulation and Supervision of Microfinance*. CGAP and World Bank, Washington, DC.

27. Clerkin, B. & Quinn, M. (2021). “Institutional agents missing in action? Management accounting at non-governmental organisations”. *Critical Perspectives on Accounting, pp. 1 – 16*

28. Cuijpers, R., Buijink, W., (2005). “Voluntary adoption of non-local GAAP in the European Union: A study of determinants and consequences”. *European Accounting Review, Vol. 14, Issue No. 3, pp. 487 – 524.*

29. Daff, L. & Parker, L.D (2021). “A conceptual model of accountants’ communication inside not-for-profit organisations”. *The British Accounting Review, Vol. 53, Issue No. 3, pp. 1 – 20*

30. Dillard, J.F., Rigsby, J.T. & Goodman, C. (2004). “The making and remaking of organizational context”, *Accounting, Auditing & Accountability Journal, Vol. 17, Issue No. 4, pp. 506 – 42.*

31. Ding, Y., Hope, O., Jeanjean, T., & Stolowy., H (2007). “Differences between domestic accounting standards and IAS: Measurement, determinants and implications”, *Journal of Accounting and Public Policy, Vol. 26, Issue No. 1, pp. 1 – 38*

32. Downes, J.F., Flagmeier, V. and Godsell, D. (2018). “Product market effects of IFRS adoption”. *Journal of Accounting and Public Policy*, Vol. *37, Issue No.* 5, pp. 376 – 410.

33. Elad, C.M. (1992). *Influences on the Harmonisation of Accounting and Disclosure in Cameroon.* PhD Thesis, Accounting and Finance, Glasgow University, UK.

34. Elad, C.M. & Tumnde, M. (2007). *Uniform Act Organizing and Harmonizing Accounting Systems in the Signatory States to the Treaty on the Harmonization of Business Law in Africa*

with Commentaries. Paris : Centre National de la Recherche Scientifique*,* CNRS-Juriscope: 1 – 46.

35. El-Gazzar, S.M., Finn, P.M., & Jacob, R. (1999). "An empirical investigation of multinational firms' compliance with International Accounting Standards". *The International Journal of Accounting, Vol. 34, Issue No. 2, pp. 239 - 248*

36. Enonchong, N. (2007). "The Harmonization of Business Law in Africa: Is Article 42 of the OHADA Treaty a Problem?" *Journal of African Law, Vol. 51 Issue No. 01, pp. 95 – 116.*

37. Ezzamel, M., Robson, K., Stapleton, P. & McLeanb, C. (2007), "Discourse and institutional change: giving accounts and accountability", *Management Accounting Research, Vol. 18, Issue. No 2, pp. 150 – 71.*

38. Firth, M., Rui, O.M., & Wu, W. (2011). "Cooking the books: Recipes and costs of falsified financial statements in China", *Journal of Corporate Finance, Vol. 17, Issue No. 2, pp. 371 – 390*

39. Flick, U. (2002), *An Introduction to Qualitative Research*, Sage Publications, Thousand Oaks, California.

40. Fordham, D.R., & Hamilton, C.W. (2019). "Accounting information technology in small businesses: An inquiry". *Journal of Information Systems, Vol. 33, Issue No. 2, pp. 63 – 75*

41. Francis, J.R., Khurana, I.K., Martin, X., Pereira, R., (2008). "The role of firm-specific incentives and country factors in explaining voluntary IAS adoptions: Evidence from private firms". *European Accounting Review, Vol. 17, Issue No. 2, pp. 331 – 360.*

42. Gaël, R., & Anand, R. (2013). Behavioural Economics and Public Sector Reform: An Accidental Experiment and Lessons from Cameroon. Policy Research Working paper 6595, World Bank, Africa Region. https://openknowledge.worldbank.org/bitstream/handle/10986/16046/WPS6595.pdf?seque [Accessed 23 September 2021]

43. García, M.D.P.R., Alejandro, K.A.C., Sáenz, A.B.M., Sánchez, H.H.G., (2017). "Does an IFRS adoption increase value relevance and earnings timeliness in Latin America?" *Emerging Markets Review, Vol. 30, 155–168.*

44. Greenwood, R. & Hinings, C.R. (1993), "Understanding strategic change: the contribution of archetypes", *The Academy of Management Journal, Vol. 36, Issue No. 5, pp. 1052 – 1081.*

45. Greenwood, R. & Hinings, C.R. (1996), "Understanding radical organizational change: bringing together the old and the new institutionalism", *Academy of Management Review, Vol. 21, Issue No. 4, pp. 1022 – 1054.*

46. Gupta, P.K. and Sharma, S. (2021). “Literature review on effect of microfinance institutions on poverty in South Asian countries and their sustainability”. *International Journal of Emerging Markets*.

47. Healy, P.M, & Palepu, K.G. (1993). “The effect of firm’s financial disclosure strategies on stock prices”. *Accounting Horizons, Vol. 7, Issue No.1, pp. 1 – 11*

48. Hopper, T., Lassou, P, & Soobaroyen, T. (2017). “Globalisation, accounting and developing countries. *Critical Perspectives on Accounting, Vol. 3, Issue No. 2017, pp. 125 – 148*

49. Hopwood, A.G., (1987). “The Archaeology of accounting systems”. *Accounting, Organisations and Society, Vol. 12, Issue No. 3, pp. 207 – 234*

50. Hopwood, A.G., (1990). “Accounting and Organisation Change”. *Accounting, Auditing & Accountability Journal, Vol. 3, Issue No. 1, pp. 7 – 17*

51. Holthausen, R.W. (2009). “Accounting Standards, Financial Reporting Outcomes, and Enforcement”. *Journal of Accounting Research, Vol. 47, Issue No. 2, pp. 447 – 458.*

52. Jones, M.J., & Mellett, H.J., (2007): “Determinants of changes in accounting practices: Accounting and the UK Health Service”. *Critical Perspectives on Accounting, Vol. 18, Issue No. 1, pp. 91 – 121*

53. Kirkpatrick, I., & Ackroyd, S. (2003) “Archetype theory and the changing professional organisation: A critique and alternative”. *Organisation, Vol. 10 Issue No. 4, pp. 731 – 750.*

54. Kurfi, B.U. (2008). Overview of credit delivery channels in Nigeria in *"The Role of Microfinance in the Economic development of Nigeria"*. Publication of Central Bank of Nigeria, Vol. 32, Issue No. 1

55. Leavy, P. (2017). *Research design: Qualitative, Quantitative, Mixed methods approaches, Arts-Based and Community – Based participatory research approaches.* The Guildford Press, London

56. Liguori, M, & Steccolini, I. (2011),"Accounting change: explaining the outcomes, interpreting the process", *Accounting, Auditing & Accountability Journal, Vol. 25 Issue No. 1, pp. 27 – 70*

57. Long, I., (2009). Perceptions of Microfinance in Cameroon: A Case Study of UNICS, Yaoundé. *Independent Study Project (ISP) Collection*. 729. Available online at: https://digitalcollections.sit.edu/isp_collection/729 [Accessed 13 September, 2021]

58. Lukka, K. (2007), “Management accounting change and stability: loosely coupled rules and routines in action”, *Management Accounting Research, Vol. 18, Issue No.1, pp. 76 – 101.*

59. Macintosh, N.N., & Scapens, R.W. (1990). “Structuration theory in management accounting”, *Accounting, Organisations and Society, Vol. 15, Issue No. 5,*

pp. 455 –477

60. Mahoney, P.G., (2009). "The Development of Securities Law in the United States." *Journal of Accounting Research, Vol. 47, Issue No. 2, pp. 325 – 347*

61. Marshall, M. N. (1996). "Sampling for qualitative research". *Family Practice, Vol.13, Issue No. 6, pp. 522 – 526*

62. Meyer, J. W. (2008). Reflections on institutional theories of organizations. *The sage handbook of organizational institutionalism*, 790-811. Avialable online at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.465.8438&rep=rep1&type=pdf [Accessed 01 November, 2021]

63. Meyer, J.W. & Rowan, B. (1977), "Institutionalised organisations: formal structures as myth and ceremony", *American Journal of Sociology, Vol. 83, Issue No. 2, pp. 340 – 63.*

64. Miles, M.B. & Huberman, A.M. (1994). *Qualitative Data Analysis: An Expanded Sourcebook of New Methods*. Sage, Thousand Oaks, California.

65. Miller, P. (1994), "Accounting as social and institutional practice: an introduction", in Hopwood, A.G. and Miller, P. (Eds), *Accounting as Social and Institutional Practice*, Cambridge University Press, Cambridge, pp. 1 – 39.

66. Miller, P. & O'Leary, L. (1987). "Accounting and the construction of the governable person", *Accounting, Organisations and Society, Vol. 12, Issue No. 3, pp. 235 – 265*

67. Moussa, B. (2010). "On The Development of West African Accounting System". *International Journal of Business and Management, Vol. 5, Issue No. 5, pp. P62 - P73*

68. Nobe, A.A., (2015). "Reforming public sector accounting and financial systems: A critique of Cameroon and Nigeria". *International Journal of Accounting and Business Finance, pp. 86 – 101.*

69. OHADA, (2011) *Acte Uniforme Relatif Au Droits des Societes Cooperatives*. Yaounde: OHADA, pp. 1 – 96

70. Palepu, K.G., Healy, P.M., Wright, S., Bradbury, M., & Coulton, J. (2020). *Business Analysis and Valuation using Financial Statements.* 3rd Edition, Cengage Learning, Australia

71. Patton, M.Q. (2002). *Qualitative research and evaluation methods*. Sage, Thousand Oaks, California.

72. Pelekh, U., Khocha, N & Holovchak, H. (2020). "Financial statements as a management tool". *Management Science Letters, Vol. 10, Issue No. 1, pp. 197 – 208.*

73. Percival, J. (2008). The 1961 Cameroon Plebiscite: Choice or Betrayal? Langaa Publishers, Mankon

74. Pignatel, I. & Tchuigoua, H.T. (2020). Microfinance institutions and International Financial Reporting Standards: An exploratory analysis. *Research in International Business and Finance, Vol. 54, Issue No. 101309, pp. 1 – 16.*

75. Powell, W. & DiMaggio, P.J. (1991). *The New Institutionalism in Organizational Analysis*, The University of Chicago Press, Chicago, IL.

76. Preston, A.M., Cooper, D.J., & Coombs, R.W. (1992). Fabricating budgets: a study of the production of management budgeting in the National Health Service. *Accounting Organisations and Society, Vol. 17, Issue No. 6, pp. 561–93*

77. Puxty, A.G., Willmott, H.C., Cooper, d., & Lowe, T. (1987). ”Modes of regulaton in advanced capitalism: locating accounting in four countries”. *Accounting, Organisations & Society, Vol. 12, Issue No. 3, pp. 273 – 291*

78. Rosenberg, R., Mwangi, P., Christen, R.P., & Nasr, M. (2003). *Microfinance Consensus Guidelines: Disclosure guidelines for financial reporting by Microfinance Institutions.* CGAP and World Bank, Washington DC.

79. Saudagaran, S.M., & Diga, J.G. (2000). “The institutional environment of Financial Reporting regulation in ASEAN”. *The International Journal of Accounting, Vol. 35, Issue No. 1, pp. 1 – 26*

80. Scott, W.R. (2001). *Institutions and Organizations*, Sage Publications, London.

81. SEEP (2009). Microfinance reporting initiative: Business Review Model. Washington, DC. https://seepnetwork.org/files/galleries/619_MFRS_Business_Models_English.pdf [Accessed: 04 November 2021]

82. Suharsono, R.S., Nirwanto, N., & Zuhroh, D. (2020). “Voluntary disclosure, financial reporting quality and asymmetry information”. *The Journal of Asian Finance, Economics and Business. Vol. 7, Issue No. 12, pp. 1185 – 1194*

83. Tsamenyi, M., Cullena, J. & Gonza´lez, J.M. (2006), “Changes in accounting and financial information system in a Spanish electricity company: a new institutional theory analysis”, *Management Accounting Research, Vol. 17, Issue No. 4, pp. 409 – 32.*

84. Tushman, M.L., Newman, W.H. & Romanelli, E. (1986). “Convergence and upheaval: managing the unsteady pace of organizational evolution”, *California Management Review, Vol. 29, Issue No. 1, pp. 29 – 44.*

85. Van Auken, Howard E.; Ascigil, Semra; Carraher, Shawn (2017). “Turkish SMEs' use of financial statements for decision making”. *The Journal of Entrepreneurial Finance. The Academy of Entrepreneurial Finance, Vol. 19, Issue No 1* https://www.econstor.eu/bitstream/10419/197540/1/1663011222.pdf [Accessed:25 November 2021]

86. Yin, R.K. (2009). *Case study research: Design and methods*, Sage Publications, Inc, Thousand Oaks

Pay Per Click Attribution: Effects on Direct Search Traffic and Purchases

Toni Raurich and Dr Joan Llonch-Andreu

Toni Raurich is with the Autonomous University of Barcelona, Spain (e-mail: toni.raurich@uab.cat).

Abstract**—**This research is focused on the relationship between Search Engine Marketing (SEM) and traditional advertising. The dominant assumption is that SEM does not help brand awareness and only does it in session as if it were the cost of manufacturing the product being sold. The study is methodologically developed using an experiment where the effects were determined to analyze the billboard effect. The research allowed the cross-linking of theoretical and empirical knowledge on digital marketing. This paper has validated that, this marketing generates retention as traditional advertising would by measuring brand awareness and its improvements. This changes the way performance and brand campaigns are distributed within marketing departments, effectively rebalancing budgets moving forward.

Keywords**—**Search Engine Marketing (SEM), Click-through ratios (CTR), Pay-per-click (PPC), Marketing attribution.

2 - Problem statement

SEM tools such as PPC traffic and its sustained causal link with the advertiser's website visits (direct traffic) and subsequent conversion represent a relevant research topic for the emerging knowledge production in advertising management contexts in the digital age [1]. Concerns about its value are the backbone of our research, with brand awareness and what concerns the value of CTR as contributing parameters in understanding digital marketing from the perspective of digital consumers during an internet search [2].

The interest in this topic is substantially based on the previous studies consulted and cited as the basis of the argumentation that has been exposed. However, advertisers have only had at their disposal information generated through descriptive studies in broad contexts, in which there is evidence of the need to undertake future research under empirical approaches supported by causal relationship designs, to obtain specific data on the performance of their digital marketing strategies in metrics such as direct traffic, the number of sessions, click-through rate and conversion rate, issues that together with the understanding of the meaning given by users to brand awareness they are repeatedly noted as matters of interest for future research.

Despite this, it is surprising that there are few empirical studies on this. The investigations reported are still incipient in terms of the methodological perspective of the empiricist-correlational-causal approach and regarding experimental or multi-method designs that include qualitative information complementary to the explanatory one, so it is necessary to develop heuristic lines that consider directly analyzing causal relationships sustained between the SEM and the referred metrics, to evaluate the "billboard effect" with greater experimental precision. In the same way, the interpretation and understanding of the narrative of the users regarding the perception of the brand. This combination of approaches ensures a more holistic view of digital marketing.

3 - General theoretical framework of the study

An organization has strategic resources that give the opportunity to develop competitive advantages over its rivals (improve results). Resources are very diverse, but their characteristics are values that must be tangible, unique, difficult to imitate and irreplaceable. This sets up the company with a sustainable competitive advantage [3].

One thing is the resources that an organization has, and another thing is the capabilities, what a company can do based on the resources it has. The dynamic capabilities are knowing how to improve and adapt to changes in the environment [4].

This resource-based approach is fully in place to address current issues. Marketing tools and new digital marketing tools represent these resources and capabilities that can give an organization a competitive advantage over its rivals.

With the advent of Web 3.0 and the Internet, companies are adapting and using digital marketing. Following the signature-focused definition of the American Marketing Association [5], digital marketing could be considered as the process of creating and communicating a product in which value is added for customers and other stakeholders, who participate in institutional activities and protocols mediated by digital technologies. To this, we add that the purpose of digital marketing is to achieve a deep knowledge and understanding of the customer to the point that the product or service sells itself by being appropriately tailored to the customer [6]. We also add the perspective that all parties involved participate in a process in which an adaptation occurs whereby organizations synergize with customers and partners in creating, communicating, delivering and sustaining the value intrinsic to the product or service. [7] And more and more companies are striving to have a stronger presence on the Internet in general and in search engines (SE) in particular. These search engines have become the strategic platform for companies [8]. They have become the main source of information for users and there are studies that demonstrate their effectiveness in directing traffic to relevant brand offerings [9].

Search platforms work with two advertising formats, these are SEM and optimization approaches (SEO). SEM works primarily on keyword-based searches and website visibility in search engines. An adequate advertising budget, but above all a concordance of keywords, play a vital role in the result [10].

The adaptation process developed by these digital technologies continues to create value in a different way in these new digital environments and builds fundamental capabilities to create that value together for your customers and the organization. These processes enabled by digital technologies create value through new customer experiences and through interactions among customers, in addition to referring to the points of contact resulting from that digital activity developed by the company [7].

To the above it is necessary to add what Christodoulides [11] said in relation to marketing and, particularly, with respect to brand value when he suggests “that emotional connection, online experience, responsive service nature, trust, and fulfillment constitute five independent, yet correlated dimensions of the construct”. This shows disagreement with those who conceptualize it only as the passive or reactive result of the marketing intervention.

Research proved that consumers move through the various channels at various stages of the purchase process, from knowledge and familiarity, to evaluation and purchase of the service or product. That is why Anderl [12] considered that "interaction effects between contacts across channel types indicate an increase in purchase propensity and thus serve as a good proxy for progress in multistage purchase decision processes". This may be linked to the fact that consumers who consistently receive value during their journey through the purchase funnel are more likely to be loyal customers, an issue that tends to be achieved when generating data obtained by analyzing the path followed from the purchase funnel; first contact until the conversion is made.

Customers can collect information from search engine-focused research and read reviews from other customers on retailer sites or third-party forums not controlled by the seller. In addition, in the digital environment, customers can move forward in their decision-making journey in fundamentally new ways.

The above approach, explored from the perspective that the consumer is co-creator of the meaning of brand value, is consistent with the arguments of Baye, De Los Santos and Wildenbeest [13] that brand awareness should be included as a strategic part of SEO, due to its beneficial effects on traffic and organic clicks, so they consider that investments in this parameter are essential. They assert that, once insured by other click drivers, “consumers tend to click retailers that are more recognized, trusted, have reputations for providing value (in terms of prices, product depth or breadth), service (well-designed websites, return policies, secure payment systems)”. This mirrors the approaches presented by Kotler and Keller [14] regarding the basic assumptions of the Zaltman Metaphor Elicitation Technique (ZMET).

Regarding this, we must add what Baker [15] reported with respect to the fact that the behavior of individuals responds to environmental stimuli and that these are processed internally to emit a response or behavior, preceded by affect and cognition. In reality, consumer behavior coincides with the dominance adjusted to the environment, whose connotative character is handled in the double of the restriction and, at the same time, the promotion of the action and behavior of the individual, since it has been exposed to marketing and brand knowledge and value has emerged.

Seen from this perspective, digital marketing has better elements of judgment that also allow to argue, for example, the billboard effect, through which indirect mechanisms encourage a brand to take advantage of collateral advertising platforms to appear before the consumer and make the sale or improve its market positioning. Regarding this, Chiou [16] states the following: evidence suggests that when third party sellers emphasize the brand name in their advertisements, they do not stand out and customers are more likely to make a purchase through the direct channel, after ignoring sponsored offers.

The above has a competitive advantage in the digital environments referred to by Kannan [7] when he states: Significant transformations in consumer behavior are predicted, derived from digital technologies and devices, such as the Internet of Things (IoT) and various products associated with artificial intelligence, as well as those related to deep learning. Therefore, the prospects are promising in terms of research that is limited to this theoretical framework, in which digital marketing, brand awareness and the billboard effect interact synergistically to serve as an argument and basis for answering the questions of research raised here. This is despite the fact that much cooperative research is still needed among academics, clients, advertising agencies and goods and services companies.

4 - Main contributions of the research and research gaps

The purpose of this study is to evaluate the sustained causal relationships between positioning marketing (SEM), as digital marketing tools, in relation to direct traffic (TD), CTR, and their causal link with results. Advertisers' product and service conversion goals and the possibility of interpreting users' perception regarding brand awareness in qualitative terms are included within the heuristic intentions. It constitutes a relevant contribution by generating updated data in a substantial and contextualized way for the analysis of attribution models [17] that allow a better understanding of the characteristics of this new synergy between technology and marketing, in the hands of the main entities involved (agencies, companies and consumers), since the main considerations for budget and investment decision-making are extracted from them [18].

The knowledge that will emerge in the terms described is in tune with the most advanced in the advertising industry. It is confirmed that this study will contribute to both academia and the advertising industry, laying the foundation for future research but also making marketing work more efficient. It will provide more empirical evidence to explain the current state of knowledge in this area. It also highlights the contributions regarding the profitability of the campaigns. In strictly academic terms, the contribution to research programs related to digital marketing is evidenced by a multiplicity of empirical and methodological information of considerable importance.

In accordance with the above, this research is framed in the proposal of a theoretical and empirical contribution conceived in the context of a heuristic line that, although it has predecessors in a broad environment, specifically requires the sum of greater efforts to address objectives of knowledge on digital marketing in quantitative and qualitative perspectives, which considers the dynamism of emerging and disruptive ecosystems in which agencies interact with their campaigns, potential clients and the product/service they offer.

5 - Literature review

Search engine marketing (SEM) emerges as a relevant issue in digital advertising and has gained greater relevance in today's cybersociety. This study considers the theoretical background on SEM and the conventional metrics described in broad contexts, but taking into account the need for future research that addresses the issue from more pragmatic perspectives. In the research used as background, the SEM is revealed as a digital marketing tool in association with conventional metrics, directly and indirectly, known as an approach to the billboard effect, a very useful information provider for strategic management.

Based on the above, the keywords that guide this study are positioning marketing, brand awareness, direct search traffic, conversion rate and click-through ratio, in addition to the billboard effect.

The internet has produced significant cost savings in many sectors of the economy, impacting its productivity. It helped lower prices for consumers, resulting in faster growth in living standards. This huge development generated high digital dividends, promoted trade, made the use of capital more efficient and has been stimulating greater competition, to the point that institutions and companies switched their efforts to access potential customers through marketing strategies consistent with the reality of a market in permanent digital evolution according to the World Bank Group [19].

Also remarkable is the growth of the online search were “the average consumer now practices, about twice per day, an activity that barely existed 15 years ago” [20]. With over four billion people choosing to shop over the internet [21]. This shift in paradigm [20] meant advertisers changed their marketing budgets from offline (Television, radio or print) to online (Facebook, mainly Google) following those consumers who prefer the less intrusive navigation inside the search engines compared to television [22]. Online marketing forced advertisers to continually reinvent themselves in order to respond to the growing competition to reach, convert and later generate loyalty in their prospective clients. Classic brands had to create new competitive advantages in order to remain perceived as innovative and aspirational, hence there is a digital transformation being built in order to keep up with the speed of innovation required to operate in this new field.

In this context, for native e-commerce entities, two techniques emerged above the rest. The first one is search engine optimization (SEO) strategies which mean increasing the amount and quality of visitors to your website through non-paid techniques, mainly through writing good quality content [13]. Second is Search Engine Marketing (SEM) which consists of an advertising model where advertisers pay every time a user clicks on one of their advertisements [23]. Google alone made $134 billion in revenue from advertisers in

SEM. The latter is the bulk of the investment and where we would like to focus especially on this paper.

Traditionally, an SEM campaign is measured in terms of investment and its direct return. Tracking tools such as Google Analytics gives the advantage to do it for free. On the other hand, television is measured in less tangible manners such as increase in brand awareness or top of mind percentage. This created a different bar of measuring, making it more difficult for SEM as it can all be measured, compared to weaker infer techniques. The reality proved that nothing is black or white, and that disconnecting investment in television to the performance on SEM is misleading as investment on one channel brings benefits to the other.

The Billboard Effect, according to Anderson [24], in the hospitality industry, refers to the phenomena where potential guests see your hotel on an online travel agency (OTA), but then decide to visit — and ultimately book through — your website directly. A notorious brand in this aspect is Booking.com, a publicly traded company in NASDAQ, which managed to grow its brand awareness through investing solely in Google with not above the line activities. All of it came out of investment in quantitative, performance-based activities, and we will try to reproduce its effects in this experiment. Our thesis is that the billboard effect can also be reproduced using Google AdWords, when our brand is exposed to customers when searching for hotel terms, hence we will use the billboard effect to refer to a brand exposition on Google AdWords and not to its traditional definition.

One of those aspects that must be constantly reviewed is the one related to attribution, approaching from a different perspective and with respect to which Berman refers to as a highly relevant metric that must be assumed as “a fully strategic choice by advertisers and publishers and not just as a measurement technique” [25]. The orientation towards strategic management, according to the author, needs to be considered in the future, so it deserves new perspectives that generate more information about the final conversion in the advertiser site.

Regarding the above, Du, Yuxing, Linli and Kenneth (2019) [26] states: "Future research needs to address the question of whether the rate of immediate online response is positively correlated with the amount of online and offline response accrued over time and, ultimately, with incremental sales attributable to a single spot”. This corroborates the interest in having specific information on the effects of advertising, such as search engine traffic or PPC on the behavior of users until purchase decisions are made in their direct traffic in the landing page.

On the other hand, Li and Kannan [27] raise other thematic aspects closely related to attribution, the budget, search and keywords, as well as the billboard effect when they express the need to “estimate the carryover and spillover effects of prior touches at both visit and purchase stages is necessary to correctly measure the incremental contribution of multiple channels and overlapping campaigns and to assist decisions on optimizing marketing budgets”. This confirms the concern to specify the value of each of the direct and indirect variables in the generation of the expected results during the planning of the advertising strategy; insisting on the urgency of giving advertisers and publishers sufficient evidence to make the most accurate decisions possible [28]. The main problem comes from the fact that Google Analytics, the main performance monitoring tool, does not understand the behavior of those who surf the web when they change devices; for example, using first a mobile and then a tablet to complete the purchase.

Kannan, Reinartz and Verhoef [29] make other contributions in relation to the research requirements, when they express: “appropriate attribution strategies can allow attribution results be used in real-time for targeting purposes. As more and more marketing interventions get automated, developments in this area are needed to ensure that such marketing campaigns result in maximal ROI”. With this, they make clear the growing interest in deepening research in these thematic fields, which aim towards obtaining first-hand data and information for immediate decision making.

6 - Key literature

The foundations of the documentary review for this research are based on the theme of positioning marketing (SEM) as a digital marketing tool in the Information, Communication and Knowledge Society (ICT), related to the importance of advertising metric data classics.

ICTs are leading to an enormous development that, according to Evans [30], is opting for the IoT and granting cyber citizenship within the framework of Web 2.0. According to the World Bank Group [19], its impact tripled and generated high digital dividends, promoted trade, made the use of capital more efficient and has been stimulating greater competition, to the point that institutions and companies channel efforts to access your potential clients through marketing strategies consistent with the reality of a market in permanent digital evolution.

Based on the above, online marketing is forced to continually reinvent itself to respond to the growing requirements presented by product and service providers to reach their prospects until they achieve conversion and even achieve loyalty. The search engine optimization (SEO) strategies and the management of paid links in search engines emerged in this context. SEM is considered by Anderson et al. [31].

This scenario leads marketing managers to rely on online metrics such as CTR and cost per click (CPC), as well as calls to action (CTA), related to conversion rate and return on investment (ROI), but with higher objective specificity for return on advertising investment (ROAS), providers of sufficient evidence to make decisions regarding advertising budgets and strategies.

However, there are opacities noted concerning some of the digital marketing tools and their classic metrics, given the inaccuracies about attribution in terms of final conversion, according to the approach of Li et al. [32], in the context of its progression through the funnel, by not considering the indirect effects of the use of other channels, in addition to the continuous interactions of a dynamic and contextual nature, raised by Kireyev, Pauwels and Gupta [33] as pending issues.

One of those aspects that must be constantly reviewed is that referring to attribution, approached from a different perspective and with respect to which Berman [34] refers to as a metric of enormous relevance that, even, has to be assumed as “a fully strategic choice by advertisers and publishers and not just as a measurement technique". The orientation towards strategic management, according to the author, needs to be considered in the future, so it merits new perspectives that generate more information on the final conversion in the advertiser's site.

7 - Key research findings

Previous studies have aimed to describe the characteristics of the metrics and even the associations between variables, but no research has been proposed to seek the empirical establishment of the sustained effects of the variables discussed here.

Regarding the above, Danaher and Heerde [35] say that "the measurement of multimedia attribution is an area where there much opportunity for marketing science principles to make an impact on practice...". This corroborates the interest in having in a concrete way information on the effects of advertising, such as the traffic of paid search engines or PPC on the behavior of the users until the purchase decisions are made in their direct traffic in the landing page.

On the other hand, Li et al. [27] raise other thematic aspects closely related to attribution, budget, search and keywords, and the billboard effect when they express the need to “… estimate the carryover and spillover effects of prior touches at both visit and purchase stages is necessary to correctly measure the incremental contribution of multiple channels and overlapping campaigns and to assist decisions on optimizing marketing budgets”. This confirms the concern to specify the value of each of the direct and indirect variables in generating the expected results during the planning of the advertising strategy. It insists on the urgency of giving advertisers and publishers sufficient elements of judgment to make the best possible decisions.

Kannan et al. [29] make other contributions concerning research requirements, when they state: “... paid search campaigns are generally automated and run on a daily basis by algorithms that determine campaign specifics. In such a context, misattributions of credit for keywords can lead to a significant drop in campaign ROIs…”. With this, they clarify the growing interest in deepening research in these thematic fields, which point towards obtaining first-hand data and information for immediate decision making.

Based on the above, some questions and proposals for future research arise that could lead to methodologies, models, and, in general, useful knowledge to improve digital marketing, establishing the SEM, brand awareness, and metrics as a reference.

Connected to the proposals that the antecedents leave for the discussion, the approaches of Kapoor, Dwivedi and Piercy [36] are rescued who affirm that knowledge about the circumstances in which causal relationships are executed can guide the discussion regarding their link not only with traffic but with conversion, respecting the company's perspective on determining attribution and the purpose of your campaign. To this must be added the growing interest in brand awareness as relevant information for decision making in terms of investment and campaigns from the hand of strategic management supported by ICT.

8 - Main hypothesis

As the first step of inferential verification, it has been considered to assume that the basis of the test is oriented to estimate the relationship between the study variables, for which the inferences that assume as probable a relationship between the positioning marketing strategy and the direct traffic; the same about conversion, click-through rate and the number of sessions on the online travel agency website.

Now, taking into account the premise that the positioning marketing strategy (SEM) implies an exposure on the internet that favors the presence of the online travel agency in cyberspace, and that this affects the considerations of the cybernaut who explores the network, then it can be inferred that the number of sessions that are opened on the web pointing towards the agency's site, the direct traffic and the conversion will be affected in some way. This circumstance is known as the billboard effect and serves as a framework for inferences under the hypothetical-deductive method applied in the methodological route described later.

A notable brand in this regard is Booking.com, a NASDAQ publicly traded company, which managed to increase its brand awareness by investing solely in Google with no above-the-line activities. It all came from investing in performance-based quantitative activities, and we will try to reproduce its effects in this experiment.

Another important aspect that must be considered in the context of the study is regarding user navigation and clicks related to the search. Jerath, Ma and Park [37] stated that "…click behavior on the search results page is governed by two components of the model: the overall propensity to click and the likelihood to search for information in the sponsored versus organic listings". This highlights the importance of accessing data referring to the route followed by the client from the beginning of the navigation until the conversion is completed.

Obtaining data on the path that a potential client follows or who has actually converted represents valuable information. In this way, marketing managers can take into account, for example, the fact that the client has reached their website after having appeared on the website of a third party that has advertised it. This is known as the billboard effect. An example of this approach is argued by Anderson [24] when he states that the theoretical basis for this phenomenon provides an explanation for understanding how the potential guest accesses information about the hotel through its listing on the OTA, but finally makes the reservation directly through the hotel's own channel or its subsidiaries. The same author argues that potential customers use the exposure of a product or service through a third party to carry out a first exploration, but they do not make the conversion in the consulted portal, but end up converting the company directly on the web.

With this, the billboard effect becomes relevant, and on that basis, they are proposed as probable and plausible solutions to the research questions. Said premises and argumentative basis are raised in the following hypotheses:

H1: It is stated that there are significant differences between the averages of sessions due to direct traffic to the advertiser's website when comparing when the Positioning Marketing Strategy (SEM) ceases and when it is active.

H2: Significant differences are expected between the proportion of clicks on the advertiser's website when comparing when SEM ceases and when active.

H3: The significant differences between the conversion rate averages are established when comparing them when the Positioning Marketing Strategy (SEM) ceases and when it is active.

But in addition, inferences will be made that establish the association of the SEM with the metric indicators of brand awareness, both in quantitative terms (established in a hypothesis) and qualitative (as categories and semantic networks). How this will be done methodologically will be explained in the next section.

9 - Research introduction

The considerations made in this research have as a reference the undeniable boom of Information and Communication Technologies (ICT) today, being the advertising field one of the most influenced. Based on this, it is assumed that the positioning marketing strategy (SEM-PPC) promotes digital exposure throughout the network, which is an influence on the prospect who navigates on the internet. It is plausible to assume that a user's sessions on the agency's website, direct traffic, and conversion could be influenced by entering that positive feedback loop.

The above, understood as the billboard effect, serves as a contextual theoretical argument that seeks in a plausible way, and with a high level of probability, to develop the research route in empirical terms. In this same sense, Kapoor et al. [36] state that causal relationships can guide the discussion regarding their link, not only with traffic, but with conversion, respecting the company's perspective, regarding the determination of the attribution and purpose of your campaign.

Based on the above, some questions arise that could lead to methodologies, models and useful knowledge for the purposes of digital marketing, establishing as reference the search traffic for payment in terms of the following: is there a sustained causal relationship between positioning marketing strategies (SEM) and direct search traffic? The concern arises from the systemic approach to the budget, with the understanding that investment means targeting efficient resource management, with which another question arises related to the cessation of SEM exposure and its impact on direct traffic: To what extent is direct traffic sustained after PPC has ceased? In other words: will there be a difference between the average number of visits (traffic) obtained on the landing page of the advertiser originating directly vs. the paid one? This, after having stopped the SEM strategy.

Answering the research questions, together with the discussion of the results, will constitute a valuable contribution to the lines of research in which this study is circumscribed, with the same tenor regarding the epistemic perspective and the theoretical-methodological body that will emerge from the heuristic intentions outlined here.

In terms of digital marketing, the perspective of this study is unprecedented in that it not only contemplates theoretical aspects, but also formulates an experimental design based on empirical data generated by the online travel agency website as the basis for statistical treatment and to contribute to a relevant analysis of the interactions between variables

The discussion generated from the analysis is undoubtedly important from an academic point of view, typical of empirical research. The importance lies in the fact that it provides them with sufficiently validated elements of judgment on the behavior of users on their portals, as well as on the effects of their digital marketing strategies on parameters such as direct traffic and conversion. This helps to ensure returns on advertising investment by leveraging real-time technologies.

10 - Research approach

a) Idea or topic

PPC as a digital positioning marketing tool under the billboard effect approach, and its sustained causal link to advertiser website visits (direct traffic) and subsequent conversion, represent a relevant research topic for the emerging knowledge production in advertising management contexts in the digital era.

b) Basic problem solved by this study

Despite the development of ICT in the organizational and advertising field, from which this knowledge arises in association with digital marketing, there is no updated research that considers the attribution of paid search traffic (SEM, PPC) and its differentiation in regards to direct traffic and conversion to the advertiser's site in that they take the long-term billboard effect for granted. These considerations have not been made to date, despite representing an important parameter to take into account for making tactical and strategic decisions associated with individual campaigns based on digital marketing.

This approach arises as a result of consultations in the academic-scientific and business literature, where it is shown that the tools derived from digital marketing are developed in parallel with various studies associated with the design and implementation of strategies aimed at heterogeneous, multilingual and omnipresent audiences. This thrust in the field of research aims to provide useful information to agencies and advertisers as provider organizations (small, medium and large), who believe in the field of corporate advertising to catapult brands of products and services in terms of driving to their websites (visits) to achieve conversion and even loyalty.

The interest in this topic is obvious judging by the previous studies consulted and cited as the basis for the argumentation that has been exposed. Online travel agencies have had at their disposal information generated through descriptive studies in very broad contexts, in which the need to undertake future research under empirical approaches based on causal relationship designs, in order to obtain specific data on the performance of their digital marketing strategies in metrics such as direct traffic, number of sessions and conversion, is repeatedly raised.

Despite this, it is surprising that there is little empirical research on the specific context regarding the strategy comparing SEM and paid search traffic in regard to direct traffic and conversion rate. The reported investigations are still incipient in terms of the methodological perspective of the empiricist-correlational-causal approach and in terms of experimental designs, so it is necessary to develop heuristic lines that consider evaluating the sustained causal relationships between the visits obtained (traffic) directly and the payment marketing strategy (SEM, PPC) as digital tools, to more accurately assess the billboard effect. In the same way, specify their impact as conventional and unconventional objective advertising metrics, as well as benchmarks to make the most of the different attribution models.

c) Justification of the investigation.

The purpose of this study is to evaluate the sustained causal relationships between positioning marketing (SEM) referring to PPC and direct traffic, as digital marketing tools in terms of their causal link with the objectives of advertiser’s conversion of products and services. It constitutes a relevant contribution by generating updated data in a substantial and contextualized way for the analysis of attribution models that allow us to better understand the characteristics of this new synergy between technology and marketing in the hands of the main entities involved (agencies, companies and consumers), since the main considerations for budgetary and investment decision making are extracted from them.

With this, it is understood that the causal interaction between SEM and direct traffic represents that source of consultation about the value given to digital tools as a traffic activator, towards real conversion, in accordance with the purposes of the campaign related to attracting prospects to landing pages.

In accordance with the above, it is inferred that this research will generate relevant contributions to evaluate the theoretical and empirical knowledge about these digital marketing strategies, causally related to conventional and unconventional metrics of particular advertising performance, to be enrolled in lines of research associated with different SEM models, in addition to characterizing the dynamics of these digital tools in cyber-ecosystems, for economic-financial management purposes in the framework of the campaigns managed by the agencies. From this heuristic, knowledge will be derived that will improve both the state of the art and the real and objective practice of these advertising resources and strategies.

From a methodological point of view, this exploration will provide a working route for the digital marketing research line, will point out virtual instruments and protocols for the collection of little exploited data so far, as well as the classic analysis procedures and the field network experimental, contributing to the ubiquitous and expeditious processing of the data generated in cybermarketing studies.

So far, research with descriptive correlational frameworks reports an association between digital marketing strategies and conventional metrics, without necessarily implying the establishment of causality with any particular metric in an online travel agency context. However, this study goes further by using an empiricist approach with an experimental design in which questions are asked about causal (cause-effect) relationships between the study variables.

In this case, research questions are answered regarding the effects between the online travel agency's digital marketing strategies (SEM-PPC) and the metrics related to conversion, number of sessions and direct traffic on its website, an issue that has not been studied in previous research.

d) Viability of the investigation.

In the era of the IOT, connectivism, and e-citizenship, research on the topic of interest is made possible largely by the e-culture present in academia and business. It is especially made possible by the diversity of free or paid tools that are available, such as instruments for data collection, processing and exchange, to which the researcher has primary access in real time, making it clear that obtaining such data as a source of information is truthful and numerous, it makes this study feasible. In addition, there is capital and heuristic and cultural experience in the area of digital marketing to implement the work route that will be designed, in order to answer the research question: what is the effect of positioning marketing (SEM, PPC) in direct and sustained search visits (traffic) to the advertiser's site? Similarly answering the question, what is the effect on conversion?

e) Novelty of the investigation.

The knowledge that will emerge in the terms described is in tune with the most advanced in the advertising industry. It is confirmed that this study will contribute to both the academy and the advertising industry, laying the foundation for future research, but also making marketing work more efficient. It will give further empirical evidence to explain the current state of knowledge in this area and the perspectives on the use of paid search traffic, its real attribution in terms of sustained effects towards direct traffic and the consequent conversion on the advertiser's site, which will improve understanding of this topic and decision making. The contributions regarding the profitability of the campaigns are also highlighted; and in strictly academic terms, the contribution to research programs related to digital marketing is evidenced with a multiplicity of empirical and methodological information of considerable importance.

Previous studies have aimed to describe the characteristics of the metrics and even the associations between variables, but no research has been proposed to seek empirical establishment of the sustained effects of the variables presented here (SEM, PPC) on indicators such as the direct traffic and their respective conversions, being able to promote campaigns based on approaches that focus more on branding and improve the experience of prospects to ensure conversion by direct navigation.

According to what has been stated, this research is part of the proposal for a theoretical and empirical contribution conceived in the context of a heuristic line that, although it has predecessors in a broad environment, in particular requires the sum of greater efforts to address objective knowledge on digital marketing in empirical fields, which considers the dynamism of emerging and disruptive ecosystems in which agencies interact with their campaigns, potential clients, and the product/service. The importance of positioning marketing (SEM) and its effects on direct traffic to the advertiser's site are assessed here.

f) Research targets:

General Objective: To assess the effect of the Search Engine Marketing Strategy (SEM-PPC) on direct search traffic and conversion provided by the online travel agency in UK cities during August, September and October 2020.

Specific targets:

* Determining the effect of the Search Engine Marketing (SEM-PPC) strategy on direct traffic when the strategy is switched on (ON) and off (OFF) by the online travel agent in UK cities during August, September and October 2020.

*To determine the effect of the SEM-PPC strategy on the conversion rate when the strategy is switched on and off in UK cities during August, September and October by the online travel agent.

*Compare the effect of the SEM-PPC strategy on direct traffic and the conversion rate when the strategy is switched on and off by the online travel agent in UK cities during August, September and October 2020.

g) Epistemological approach/paradigm or perspective:

This research is established with a realist ontological position regarding the nature of the object of study. But it also assumes that this reality of interest is known objectively, through the observation of quantifiable and measurable facts. Therefore, the epistemological approach assumed here is empiricism.

This study theoretically assumes that the facts of interest are observable, measurable and quantifiable, with results that can be repeated and verified. It is therefore an empiricist explanatory research from the point of view of its epistemological approach. Therefore, it can be said that the paradigm of this research is quantitative and explanatory.

h) Method/ methodology/ design/ analysis techniques/ statistics

The realist ontology, the empiricist epistemological approach and the quantitative paradigm assumed for this study lead to taking a route of the research process of verification that attends to the hypothetical-deductive method, with an experimental design, necessary to respond to the research question referring to the cause-effect relationship between the variables. In view of the above, the following is established:

- Objectivist/inductivist method.

- Quantitative methodology.

- Experimental design.

- Analysis techniques: parametric inferential statistics.

- **Inferential Statistics**: Student "t" test for independent samples, by determining the effect of the use of search engine marketing strategies (SEM-PPC), on and off, on direct search traffic to the advertiser's website and the conversion rate.

i) Variable systems:

Independent variable:

Positioning Marketing Strategy (SEM-PPC), on or off.

Dependent variables:

- Number of sessions obtained by traffic or paid search.

- Number of sessions obtained by traffic or direct search.

- Conversion rate for paid search.

- Conversion rate for direct search.

j) Hypothesis system:

As a first step of inferential testing, it has been considered to assume that the basis of the test is oriented towards dismissing the relationship between the study variables, until the data proves otherwise, so inferences are drawn that assume that a relationship between the positioning marketing strategy and direct traffic is unlikely; the same with respect to conversion and the number of sessions on the company's website. This premise and argumentative basis were stated in the following terms:

- Null hypothesis

Null hypothesis (Nh): There are no significant differences between session averages for direct traffic to the advertiser's website, when comparing when the Positioning Marketing Strategy (SEM-PPC) ceases and when it is active (Ẍ1= Ẍ2).

Null hypothesis A (Nha): There are no significant differences between the average number of sessions per paid traffic to the advertiser's website, when comparing when the Positioning Marketing Strategy (SEM-PPC) ceases and when it is active (Ẍ1= Ẍ2).

Null hypothesis B (Nhb): There are no significant differences between the conversion rate averages by paid search, when comparing them when the Positioning Marketing Strategy (SEM-PPC) ceases and when it is active (Ẍ1= Ẍ2).

Null hypothesis C (Nhc): There are no significant differences between the conversion rate averages by direct search, when comparing them when the Positioning Marketing Strategy (SEM-PPC) ceases and when it is active (Ẍ1= Ẍ2).

- Alternative hypothesis

Alternative hypothesis (Ah): There are significant differences between session averages for direct traffic to the advertiser's website, when comparing when the Positioning Marketing Strategy (SEM-PPC) ceases and when it is active (Ẍ1≠ Ẍ2).

Now, taking into account the premise that the positioning marketing strategy (SEM-PPC) implies an internet exposure that favors the presence of the online travel agency in cyberspace, and that this has an impact on the considerations of the cybernaut exploring the net, then, it can be inferred that the number of sessions that are opened on the web pointing towards the agency's site, the direct traffic and the conversion will be affected in some way.

With this, the billboard effect becomes relevant and, on that basis, the following inferences are raised as probable and plausible solutions to the research question:

Alternative hypothesis A (Aha): There are significant differences between the average number of sessions per paid traffic to the advertiser's website, when comparing when the Positioning Marketing Strategy (SEM-PPC) ceases and when it is active (Ẍ1≠ Ẍ2).

Alternative hypothesis B (Ahb): There are significant differences between the conversion rate averages by paid search, when comparing them when the Positioning Marketing Strategy (SEM-PPC) ceases and when it is active (Ẍ1≠ Ẍ2).

Alternative hypothesis C (Ahc): There are significant differences between the conversion rate averages by direct search, when comparing them when the Positioning Marketing Strategy (SEM-PPC) ceases and when it is active (Ẍ1≠ Ẍ2).

Decision rule: If $p > 0.05$ Not Rejected; If $p < 0.05$ If rejected Nh; The following should be taken into consideration: 0.05 = 5% significance or risk level.

k) Expected results

The aim is to establish the effect of positioning marketing (SEM-PPC) on direct traffic to the advertiser's site and the respective conversion of the prospect, contributing with empirical evidence for the management of Marketing 2.0, and providing tools in increasingly favorable profitability conditions for organizations, agencies and the public. From the results, the debate and the conclusions, a theoretical and empirical construction will emerge that will improve the organizational and budgetary management of digital marketing campaigns, as well as promote academic activity and research in the field of digital advertising. The aim is to establish the effect of the use of positioning marketing strategies (PPC, SEM) on direct search traffic to the advertiser's website and conversion.

In relation to conversion, it will be assessed on the basis of the sessions and income reported through the tool https://analytics.google.com/ which provides the data required for the statistical analysis process generated from the sample of cities defined for the field phase of the methodological phase.

11 - Results analysis

Once the data were obtained with the help of the Google Analytics tool from a real population, the sample of UK cities was selected, as these were the locations where the online travel agency was able to turn off its marketing strategy. This shutdown was carried out during the last week of August and part of September of the year 2020. After this period, the strategy for the sample under study was switched on again, from the last week of September to part of October, with 30 data for each condition of the strategy, ON-OFF. The data were processed by means of inferential statistics using SPSS Statistics software version 25.0.

Below there are two tables of results and their respective interpretations. Table 1 explains the statistics of the variables, number of sessions and conversion rate for paid and direct search, according to the positioning marketing strategies (SEM-PPC) in on-off condition, provided by the online travel agency during the months of August, September and October 2020. Table 2 shows the results of the student "t" test for independent samples on the variables: number of paid search sessions and direct search; and the conversion rate for paid search and direct search when the positioning marketing strategy (SEM-PPC) is on and off.

Table 1. Correlation variables, sessions and conversion rate

Condition of strategies ON-OFF		N	Average	Deviation	Average Error Deviation
Number of Sessions Paid Search	Strategy ON	30	1393.00	484.440	88.446
	Strategy OFF	30	269.37	68.200	12.452
Number of Sessions Direct Search	Strategy ON	30	80.87	24.829	4.533
	Strategy OFF	30	32.63	9.205	1.681
Conversion rate Paid Search	Strategy ON	30	2.1657	0.62839	0.11473
	Strategy OFF	30	1.5843	0.88276	0.16117
Conversion rate Direct Search	Strategy ON	30	1.3163	1.57948	0.28837
	Strategy OFF	30	1.3950	2.03577	0.37168

Source: Own elaboration.

As can be seen in Table 1, the averages obtained for the variables number of paid search sessions, number of direct search sessions and paid search conversion rate are higher when the positioning marketing strategy (SEM-PPC) is activated or turned on than when the said marketing strategy is deactivated or turned off. In the case of the variable conversion rate for direct search, the averages obtained are similar in both cases (on-off).

Table 2. Paid and direct search and conversion rate

		Levine's Test - Equality of Variances		Two-Sample t-Test for Equal Means						
									95% confidence interval of the difference	
		F	Sig.	t	df	Next (bilate - ral)	Mean difference	Standard error difference	Lower	Upper
Number of sessions. Paid traffic. ON/OFF	Equal variances	39.487	0.000	12.580	58	0.000	1123.633	89.318	944.843	1302.424
	Different variance			12.580	30.149	0.000	1123.633	89.318	941.259	1306.008
Revenue. Paid traffic. ON/OFF	Equal variances	19.062	0.000	9.977	58	0.000	48.233	4.835	38.556	57.911
	Different variance			9.977	36.823	0.000	48.233	4.835	38.436	58.031
Sessions. Direct traffic. ON/OFF	Equal variances	3.111	0.083	2.938	58	0.005	0.58133	0.19783	0.18533	0.97734
	Different variance			2.938	52.385	0.005	0.58133	0.19783	0.18442	0.97825
Revenue. Direct traffic ON/OFF	Equal variances	2.897	0.094	-0.167	58	0.868	-0.07867	0.47043	-1.02033	0.86300
	Different variance			-0.167	54.628	0.868	-0.07867	0.47043	-1.02157	0.86424

Source: Own elaboration

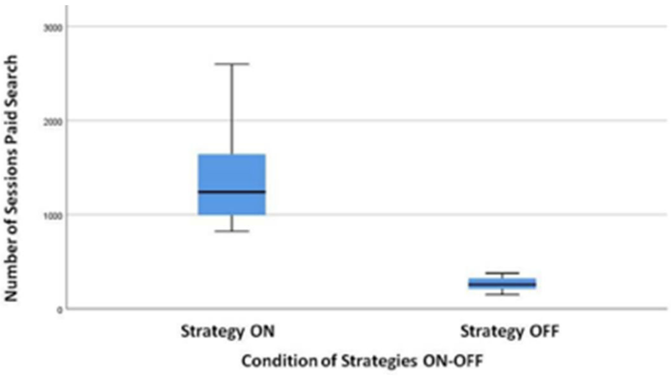


Fig. 1. Average number of sessions per search paid when the positioning marketing strategy (SEM-PPC) is on and off

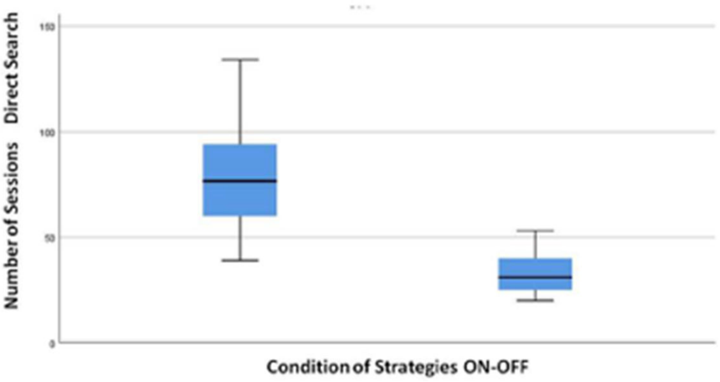


Fig. 2. Average number of sessions per direct search when the positioning marketing strategy (SEM-PPC) is on and off

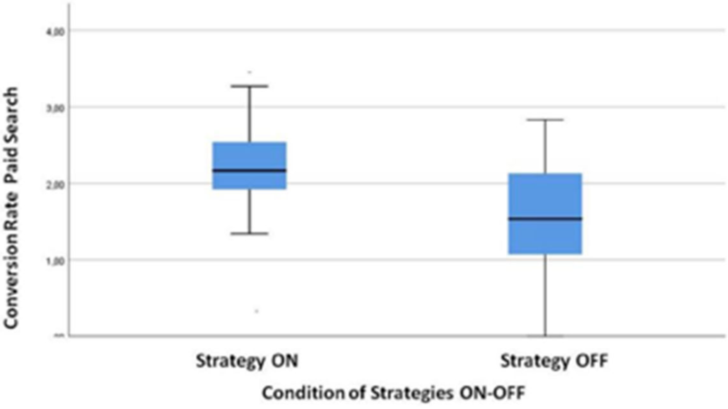


Fig. 3. Average conversion rate per search paid when search engine marketing strategy (SEM-PPC) is on and off

The results obtained for the student "t" parametric test, shown in Table 2, allow to infer the behavior of the study variables and to interpret them from the hypotheses raised.

In the case of the variables number of sessions per paid search and direct search, as well as in the case of the conversion rate for paid search, a value of p = 0.000 was obtained being lower than 0.05 (< 0.05), which allows to decide that Ho, H0a and H0b are rejected hypothesis. Consequently, it is concluded that there is sufficient evidence to establish that there are statistically significant differences between the average sessions for paid search traffic and direct search to the advertiser's website, when comparing when the positioning marketing strategy (SEM-PPC) ceases and when it is active. The same occurs with the conversion rate; in this case it only occurs when the positioning marketing strategy (SEM-PPC) is active.

This result confirms the average values observed in the case of paid search sessions when the strategy is activated (ON), which is 1393.00 against a value of 269.37 when the strategy is deactivated (OFF), clearly observing how the value is greater when the positioning marketing strategy is active (SEM-PPC).

In the case of the direct search sessions when the strategy is activated (ON), the average value obtained is 80.87 against a value of 32.63 when the strategy is deactivated (OFF), clearly observing how the value is greater when the positioning marketing strategy is active (SEM-PPC).

Based on this, we highlight the values obtained seen on figures 1 and 2, regarding the number of paid and direct sessions. We also show in figure 3 the conversion in paid traffic, and how it works against the direct traffic in figure 4 where we infer that the positioning marketing strategy (SEM-PPC) is statistically significant when it is on, preferably in the average of sessions for traffic or paid search, since it is there where profitability is obtained. This means that this strategy is effective in digital marketing.

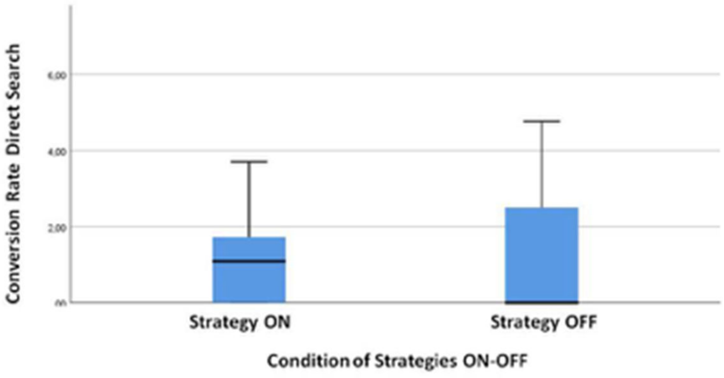


Fig. 4. Average direct search conversion rate when the positioning marketing strategy (SEM-PPC) is on (ON) and off (OFF)

In the case of the variable conversion rate by direct search, it can be seen in Table 2 that a value of p = 0.868 was obtained which is greater than 0.05 (> 0.05), which allows us to decide that for this case the H0c is not rejected. Consequently, it is concluded that there is sufficient evidence to establish that there are no statistically significant differences in the average conversion rate by direct search to the advertiser's website, when comparing the period when the positioning marketing strategy (SEM-PPC) ceases and when it is active.

When observing the average values indicated in Table 1, in the case of the direct search conversion rate when the strategy is activated (ON), the report is 1.3163 against a value of 1.3950 when the strategy is deactivated (OFF), clearly observing that the values are similar for the positioning marketing strategy (SEM-PPC) both in the on and off condition. These appreciations can be seen in a more illustrative way in Fig. 4.

12 - Conclusions, Limitations and Future Research

In this article, based on the analysis of the data obtained, a discussion is generated for the derivation of conclusions in relation to the objectives set out in the research. The effect of the positioning marketing strategy (SEM) on direct traffic is evaluated when said strategy is activated (ON) and deactivated (OFF). over the period of one quarter in the United Kingdom.

For the study, two empirical scenarios were considered, in the context of the UK, in which the sample referring to the number of direct and paid traffic sessions was measured with the SEM activated, while on the other hand it was done with this strategy disabled. Similarly, it was established in relation to the measurement of the conversion variable.

Based on the above, the evidence obtained reveals that there is a positive trend that suggests statistically significant differences between the average sessions for traffic-paid search and direct search to the advertiser's website, with paid search being favored, which is confirmed. when comparing the averages obtained in both sessions. We conclude that paid search is favored by the SEM strategy, underlining the importance of such empirical evidence as valid elements of judgement for the comprehensive explanation of the processes addressed here in specific terms and in a broad sense.

When carefully observing the results obtained regarding the effect of the positioning marketing strategy (SEM) on the conversion rate (CR), a marked increase in the latter for paid search is evident, as long as the strategy is activated (ON). This is an indication of the effectiveness of SEM in terms of this ROI metric. In this sense, it is clear that a high percentage of users who use the advertiser's website convert their intention, expressed during internet browsing, into a specific action, thereby specifying the conversion, which is inferred to improve the relationship between profits obtained and the investment made.

Based on the data generated from the United Kingdom we validate that digital marketing strategies strengthen the link between users and supplying organizations, especially when the use of digital media increases progressively and, in this specific case, the brand takes advantage of this to focus on the positioning of its product and service to meet the objectives it has set itself.

As in any empirical analysis, some limitations emerged during the research that intervened forcefully in the development of the study. The most relevant of these was the restrictions imposed worldwide as a consequence of the COVID-19 pandemic, since the study protocol coincided with its development. The pandemic significantly disrupted the daily dynamics of tourism activity, given the restrictions imposed by governments to deal with the health situation. This was reflected in the interactions of potential users with the websites being monitored, slowing down the processes involved. However, we believe that this situation could serve as a reference for a future line of research to compare pre- and post-pandemic periods in order to understand similar situations.

An additional limitation was related to the fact that we only had access to data referring to the United Kingdom, due to the fact that operationally it was not possible to deactivate SEM in more than one country simultaneously because of the losses that this would represent. As a result, the nature of the data limited our ability to obtain empirical findings on the variables in a European context and with respect to other continents. However, this should be seen as a benchmark for studies that open up the range of options to cover other countries in similar experimental circumstances in the future.

At the end of this study, beyond the results and conclusions reached here and that account for relevant contributions to the advertising industry and companies in general, it is imperative to reflect on the importance of delving into the future on the approach of similar investigations, considering samples in broader areas in terms of geographical location, with respect to the different countries and their cities in a global spectrum. It would be very useful to compare, for example, the conversion rate in relation to the type of search (paid or direct) and the on-off of a certain marketing strategy, but considering, by conglomerates, how much effect the users of countries have on different continents. The results revealed regarding the behavior of the subjects under these conditions would allow the advertiser to plan and invest in a safer and more reliable way in other latitudes.

The research also produced key findings to reach theoretical approaches related to positional marketing and the various metric parameters that explain user behavior in diverse contextual frameworks, beyond those addressed in the specific context of this study.

With the above it is clear that the theoretical reticulation is strengthened with each of the elements of judgment that emerge from this study and that bring us closer to making decisions appropriate to the demands of organizations in the context of the new post-pandemic world order.

Future research should point towards taking advantage of the findings obtained here, since they represent a step forward in terms of the explanatory knowledge of positioning marketing (SEM) on direct traffic and regarding conversion, due to the which is required to expand the data in empirical terms to improve the basis of discussion and its repercussions on digital marketing.

References

[1] Bui, M. T., Jeng, D. J. F., and Lin, C. (2015). The importance of attribution. Connecting online travel communities with online travel agents. Cornell Hospitality Quarterly, 56(3), 285-297.

[2] De Haan, Evert; Wiesel, Thorsten; Pauwels, Koen. (2016) The effectiveness of different forms of online advertising for purchase conversion in a multiple-channel attribution framework International Journal of Research in Marketing, Vol. 33 Issue 3, p491-507.

[3] Barney, J. (1991). Firm resources and sustained competitive advantage. Journal of management, 17(1), 99-120.

[4] Teece, D. J. (2018). Business models and dynamic capabilities. Long range planning, 51(1), 40-49.

[5] American Marketing Association (2021), Definitions of Marketing, https://www.ama.org/the-definition-of-marketing-what-is-marketing/

[6] Drucker, P. F. (1973). The performance gap in management science: Reasons and remedies. Organizational dynamics, 2(2), 19-29.

[7] Kannan, P. K. (2017). Digital marketing: A framework, review and research agenda. International Journal of Research in Marketing, 34(1), 22-45

[8] Rangaswamy, A., Giles, C. L., and Seres, S. (2009). A strategic perspective on search engines: Thought candies for practitioners and researchers. Journal of Interactive Marketing, 23(1), 49.

[9] Jansen, B. J., and Molina, P. R. (2006). The effectiveness of Web search engines for retrieving relevant ecommerce links. Information Processing and Management, 42(4), 1075-1098.

[10] Olbrich, R., and Schultz, C. D. (2014). Multichannel advertising: Does print advertising affect search engine advertising?. European Journal of Marketing.

[11] Christodoulides, George; de Chernatony, Leslie; Furrer, Olivier; Shiu, Eric and Abimbola, Temi. Journal of Marketing Management. Sep2006, Vol. 22 Issue 7-8, p.814. 27p. 1 Diagram, 1 Chart.

[12] Anderl, E. (20. November 2014). Three Essays on Analyzing and Managing Online Consumer Behavior. [Dissertation an der Wirtschaftswissenschaftlichen Fakultät].Universität Passau. Germany.

[13] Baye, M. R., De los Santos, B., and Wildenbeest, M. R. (2016). Search engine optimization: what drives organic traffic to retail sites?. Journal of Economics and Management Strategy, 25(1), 6-31.

[14] Kotler, P., and Keller, K. L. (2016). A framework for marketing management (p. 352). Boston, MA: Pearson.

[16] Bakker, I., Van der Voordt, T., Vink, P., and De Boon, J. (2014). Pleasure, arousal, dominance: Mehrabian and Russell revisited. Current Psychology, 33(3), 405-421.

[16] Chiou, L., and Tucker, C. (2012). How does the use of trademarks by third-party sellers affect online search?. Marketing Science, 31(5), 819-837.

[17] Barajas, J., Akella, R., Holtan, M., and Flores, A. (2016). Experimental designs and estimation for online display advertising attribution in marketplaces. Marketing Science, 35(3), 465-483.

[18] Li, H., Kannan, P. K., Viswanathan, S., and Pani, A. (2016). Attribution strategies and return on keyword investment in paid search advertising. Marketing Science, 35(6), 831-848.

[19] World Bank Group. (2016). World development report 2016: digital dividends. World Bank Publications.

[20] Joo, M., Wilbur, K. C., Cowgill, B., & Zhu, Y. (2014). Television advertising and online search.

[21] World Bank Group (2021). International Telecommunication Union, World Telecommunication/ICT Development Report and database, and World Bank Estimates. laman web: http://data. worldbank. org/indicator/IT .NET. USER. P, 2.

[22] Joo, M., Wilbur, K. C., and Zhu, Y. (2016). Effects of TV advertising on keyword search. International Journal of Research in Marketing, 33(3), 508-523.

[23] Anderson, C. K., and Cheng, M. (2017). Multi-click attribution in sponsored search advertising: An empirical study in the hospitality industry. Cornell Hospitality Quarterly, 58(3), 253-262.

[24] Anderson, C. (2009). The billboard effect: Online travel agent impact on non-OTA reservation volume. Cornell Hospitality Report, Vol. 9, No. 16.

[25] Berman, R. (2018). Beyond the last touch: Attribution in online advertising. Marketing Science, 37(5), 771-792.

[26] Du, Rex Yuxing; Xu, Linli; Wilbur, Kenneth C. (2019) Immediate Responses of Online Brand Search and Price Search to TV Ads Journal of Marketing, Vol. 83 Issue 4, p81-100.

[27] Li, H., and Kannan, P. K. (2014). Attributing conversions in a multichannel online marketing environment: An empirical model and a field experiment. Journal of Marketing Research, 51(1), 40-56.

[28] Lu, S., and Yang, S. (2017). Investigating the spillover effect of keyword market entry in sponsored search advertising. Marketing Science 36(6):976-998.

[29] Kannan, P.K.; Reinartz, Werner and Verhoef, Peter C.(2016) The path to purchase and attribution modeling: Introduction to special section. International Journal of Research in Marketing, Vol. 33 Issue 3, p449-456.

[30] Evans, D. (2011). The Internet of Things, Cisco. http://blogs. cisco. com/news/the-internet-of-things-infographic/.

[31] Anderson, C. K., and Cheng, M. (2017). Multi-click attribution in sponsored search advertising: An empirical study in the hospitality industry. Cornell Hospitality Quarterly, 58(3), 253-262.

[32] Li, H., Kannan, P. K., Viswanathan, S., and Pani, A. (2016). Attribution strategies and return on keyword investment in paid search advertising. Marketing Science, 35(6), 831-848.

[33] Kireyev, P., Pauwels, K., and Gupta, S. (2016). Do display ads influence search? Attribution and dynamics in online advertising. International Journal of Research in Marketing, 33(3), 475-490.

[34] Berman, R. (2018). Beyond the last touch: Attribution in online advertising. Marketing Science, 37(5), 771-792.

[35] Danaher, P. J., and Van Heerde, H. J. (2018). Delusion in attribution: Caveats in using attribution for multimedia budget allocation. Journal of Marketing Research, 55(5), 667-685.

[36] Kapoor, K. K., Dwivedi, Y. K., and Piercy, N. C. (2016). Pay-per-click advertising: A literature review. The Marketing Review, 16(2), 183-202.

[37] Jerath, K., Ma, L., & Park, Y.-H. (2014). Consumer Click Behavior at a Search Engine: The Role of Keyword Popularity. Journal of Marketing Research, 51(4), 480–486. https://doi.org/10.1509/jmr.13.0099.

Authors

Professor **Joan Llonch-Andreu** is a professor of marketing at the UAB, where he obtained his PhD in Economics and Business and a Diploma in Market Research. He also holds an MBA by the IMD (Lausanne, Switzerland).

Toni Raurich is a PhD candidate, who studied a Bachelor Administration and a Bachelor in communication. He completed a master in research and an MBA by the IESE Business School.

Establish Student Support Strategies for Virtual Learning in Learning Management System Based oOn Grounded Theory

Farhad Shafiepour Motlagh
Associate professor, Mahallat Branch, Islamic Azad University, Mahallat,Iran

Corresponding. Email. Farhad_shafiepoor@yahoo.com

Narges Salehi,*MA*
Dehaghan Branch,Islamic Azad University, Dehaghan,Iran

Email: narges.salehi1987@ yahoo.com

Abstract

Purpose:The purpose of this study was to support student strategies for virtual learning in the learning management system.

Methodologhy of this research:The research method was based on Grounded theory. The statistical population included all the articles of the ten years 2022-2010 and the sampling method was purposeful to the extent of theoretical saturation (n=31). Data collection was done by referring to the authoritative scientific databases of Emerald, Springer, Elsevier, Google Scholar, Sage Publication and Science Direct. For data analysis, open coding, axial coding and selective coding were used.

Results: the results showed that causal conditions include Cognitive empowerment (Comprehension, analysis, composition).

Emotional empowerment (Learning motivation, involvement in the learning system, enthusiasm for learning).

Psychomotor empowerment (Learning to master, internalizing learning skills, creativity in learning),

Conclusion: Supporting students requires their empowerment in three dimensions: cognitive, emotional empowerment and psychomotor empowerment. In such a way that by introducing them to enter the learning management system, the capacities of the system, the toolkit of learning in the system, improve the motivation to learn in them, and in such a case, by learning more in the learning management system, they will reach mastery learning.

Keyword

Student support, virtual education, learning management system

Mental Contrasting with Implementation Intentions: A Metacognitive Strategy on Educational Context

Paulino, P[ab]., Matias, A[a]., & Veiga Simão, A[a]

[a] Center for Research on Psychological Science, Faculty of Psychology of the University of Lisbon, Portugal

[b] School of Psychology and Life Sciences, Lusófona University, Lisbon

Abstract

Self-regulated learning (SRL) directs students in analyzing proposed tasks, setting goals and designing plans to achieve those goals. The literature has suggested a metacognitive strategy for goal attainment known as Mental Contrasting with Implementation Intentions (MCII). This strategy involves Mental Contrasting (MC), in which a significant goal and an obstacle are identified, and Implementation Intentions (II), in which an "if... then…" plan is conceived and operationalized to overcome that obstacle. The present study proposes to assess the MCII process and whether it promotes students' commitment towards learning goals during school tasks in sciences subjects. In this investigation, we intended to study the MCII strategy in a systemic context of the classroom. Fifty-six students from middle school and secondary education attending a public school in Lisbon (Portugal) participated in the study. The MCII strategy was explicitly taught in a procedure that included metacognitive modeling, guided practice and autonomous practice of strategy. A mental contrast between a goal they wanted to achieve and a possible obstacle to achieving that desire was instructed, and then the formulation of plans in order to overcome the obstacle identified previously. The preliminary results suggest that the MCII metacognitive strategy, applied to the school context, lead to more sophisticated reflections, the promotion of learning goals and the elaboration of more complex and specific self-regulated plans. Further, students achieve better results on school tests and worksheets after strategy practice. This study presents important implications since the MCII has been related to improved outcomes and increased attendance. Additionally, MCII seems to be an innovative process that captures students' efforts to learn and enhances self-efficacy beliefs during learning tasks.

Keyword: implementation intentions, learning goals, mental contrasting, metacognitive strategy, self-regulated learning

Teaching Students Empathy: Justifying Diverse and Inclusive Texts

Jennifer Wallbrown

***Abstract*—** It's not uncommon in the US to see news article headlines about public school teachers being scrutinized for what they are teaching or see the general public weighing in on whether or not they think certain controversial subjects should be addressed in the classroom- such as LGBTQ+ or multicultural literature. Even though this is a subject that has been written about and discussed for years, it continues to be a relevant topic in education as it continues to be a struggle to implement more diverse texts. Although it is valid for teachers to fear controversy when they attempt to create a more diverse or inclusive curriculum, it is a fight worth fighting because of the benefits students can gain from being exposed to a wide range of texts. This paper is different from others of its kind because it addresses many of the counterarguments often made to implementing LGBTQ+ or multicultural literature in secondary classrooms. It not only encourages educators to try to include more diverse texts, but it gives them the tools to address common concerns and be sound in their reasoning for choosing these texts. This can be of interest to those educators who are not English teachers because a truly diverse and inclusive curriculum would include other subjects as well- including history, art, and more. By the end of my proposed paper, readers will feel encouraged to choose more diverse and inclusive texts for their classrooms. They can also be confident that if met with opposition or controversy, as is sometimes common when implementing new texts, that they have sound arguments and reasoning for why they chose to include these texts. This reasoning is that, based on the research, studies have found there are benefits to students studying texts about those different from themselves, because it teaches them empathy and helps fight prejudice.

***Keywords*—** education, diverse, inclusive, multicultural, lgbtq+, pedagogy.

Jennifer Wallbrown is with the Marshall University, United States (e-mail: wallbrown3@marshall.edu).

Effects of External Body Movement on Visual Attentional Performance in Children with ADHD

Hung-Yu Lin

***Abstract*—** Background: Parts of researchers assert that external hyperactivity behaviors of ADHD children interfere with their abilities to perform internal cognitive tasks; however, there are still other researchers hold the opposite viewpoint, the external high level of activity may serve as the role of improving internal executive function.Objectives: Thisstudy explored the effects of external motor behavior of ADHD on internal visual attentional performance. Methods: A randomized, two-period crossover design was used in this study, a total of 80 children (aged 6-12) were recruited in this study. 40participants have received ADHD diagnosis, and others are children with typically developing. These children were measured through the visual edition of TOVA (The Test of Variables of Attention) when they wore actigraphy, their testing behavior and movement data werecollected through closely observation and the actigraphies under different research conditions. Result: According to the research result, the author found (1) Higherfrequencyof movement under attentional testing condition was found in children with ADHD, comparing to children with typically developing, and (2) Higher frequency of foot movement showed better attentional performance of the visual attentional test in children with ADHD. However, these results were not showed in children with typically developing. Conclusions: The findings support the functional working memory model, which advocated that a positive relation between gross motor activity and attentional performance within the context of attentive behavior in children with ADHD.

***Keywords*—** ADHD, movement, visual attention, children.

Hung-Yu Lin is with the Asia University, Taiwan (e-mail: otrlin@gmail.com).

Communication – A Basic Right to Everyone: How to Protect the Right of Disabled People to Communicate about Sexuality

Fanni Kevätniemi, Henna Kekkonen

***Abstract*—** When talked about sexual health, communicating about sexuality can’t be ignored. But usually it means speaking of it, and that’s a major ableist point of view. People who use AAC (augmentative and alternative communication) methods to supplement or replace speech are more than likely to be forgotten. Our accomplishment has been to create accessibility to communicating about sexuality for all individuals. By that, we can increase sexual health and well-being, raise awareness that people with disabilities are sexual beings, and fulfill their basic human right to communicate.

***Keywords*—** sexual health, picture communication, AAC, reproductive health, sex education, safety skills.

Fanni Kevätniemi is with the Metropolia University of Applied Sciences, Finland (e-mail: fanni@selkoseks.fi).

Learning Difficulties of Children with Disabilities

Chalise Kiran

Abstract**—**The learning difficulties of children with disabilities are always a matter of concern when we talk about educational needs and quality education of children with disabilities. This paper is the outcome of the review of the literatures based on the literatures of educational needs and learning difficulties of children with disabilities. For the paper, different studies written on children with disabilities and their education were collected through search engines. The literatures put together were analyzed from the angle of learning difficulties faced by children with disabilities and the same were used as a precursor to arrive at the findings on the learning of the children. The analysis showed that children with disabilities face learning difficulties. The reasons for these difficulties could be attributed to factors in terms of authority, structure, school environment and behaviors of teachers and parents and the society as a whole.

Keywords**—**children with disabilities, learning difficulties, education, disabled children

INTRODUCTION

Disability is a worldwide phenomenon that can prevail in any class, ethnicity, caste, race, gender, community, place, economic status etc. Disability can either be inborn or the result of some accidents that an individual faces in the course of his/her life. Disability is a complex multidimensional condition and poses a number of challenges for measurement [1]. Operational measure of disability varies according to the purpose and application of the data and the aspect of disability examined. Due to this, mainstreaming disability is considered difficult and not prioritized by the implementers [2].

The factsheet compiled by UNDP (2021) on persons with disabilities seem alarming. WHO indicates that around 15 per cent of the world's population, or estimated 1 billion people, live with disabilities. The people with disabilities are in largest minority. The UN development program reveals that around 80% people in developing countries have one or the other forms of disabilities. Besides, the lower educational attained countries have larger no. of persons with disabilities. In average, 11 percent better educated and 19 percent less educated have disabilities in the world. According to the data of the World Bank, 20% deprived people have some kinds of disabilities, which is largely with the disadvantaged ones. Besides, disability is common with 30% street youths [3].

Comparative studies on disability legislation show that only 45 countries have anti-discrimination and other disability-specific laws in the world. UNICEF, 2012 estimation suggests that there are at least 93 million children with disabilities in the world, but numbers could be much higher. They are often likely to be among the poorest members of the population [4].

In Nepal, the prevalence percentage of people with disabilities is estimated to be around 3.6 percent. Male and female disability rates are 4.2 percent and 3.0 percent, respectively. Physical disabilities account for 29.2 percent of people with disabilities, 22.3 percent for visual disabilities, 23.4 percent for hearing disabilities, 2.4 percent for vision/hearing disabilities, 8.6 percent for speech disabilities, 6.8% for mental retardation, and 7.3 percent for multiple disabilities[5] . However, around 2% (1.93 percent) of the population (513,321 in total) is stated to have a disability. Physical disability accounts for 36.3 percent of the disabled population, with Blindness/low Vision (18.5 percent), Deaf/hard of hearing (15.4 percent), Speech problem (11.5 percent), Multiple Disability (7.5 percent), Mental Disability (6%), Intellectual Disability (2.9 percent), and Deaf-Blindness (1.8 percent) following closely behind [6]. Another data of *Disability Atlas Nepal, 2016* reveals that 1.94% of population in Nepal is disabled and for every category, males are inappropriately more disabled in number and percentage compared to female population. There is a big difference in prevalence of rural and urban disability. The highest difference is for deaf/hard of hearing (11 times more in rural area) [7].

According to Open Society Foundations, there is no public education system in the world that is entirely free of unequal educational opportunities. There are important cross-country variations in the forms, extent, degree of systematic inequalities of educational opportunity, and the discrepancies in the policy responses. The idea of inclusive education is widely used to ensure such inconsistencies. Actually, inclusive education is to comfort the challenges of uneven education chances in education systems especially for children with disabilities [8].

When we talk about the education or learning opportunities for the children with disabilities (CWDs), there seems indeed a challenge in generic thinking. The challenges are the learning difficulties with the children which are in most of the cases created by the society, community, parents, teachers and even the schools. At least 75% of the projected 5.1 million disabled children in Eastern and Central Europe and Central Asia are denied access to a high-quality, inclusive education [9]. In underdeveloped nations, 90% of children with impairments do not attend school [10].

According to it, a study conducted in Zimbabwe found that the majority of present classrooms need to be modified to accommodate the needs of students with hearing impairment. Both heads and teachers agreed that students with hearing impairments, particularly those who are seriously afflicted, deserve specific educational facilities [11].

Significant studies in the Indian context reported the extreme exclusion of children with disabilities from the basic

Chalise Kiran is PhD Scholar with the Kathmandu University, Hattiban, Lalitpur, Nepal (e-mail: kchalise@gmail.com).

early childhood provision of health, nutrition, and preschool education. These studies reported that a large proportion of young children with disabilities had no access to early childhood education and those who had attended early education centers were denied meaningful interventions at the school level. The latest data from Census 2011 showed that out of 6.57 million people with disabilities in the age group 5-19 years, 1.75 million (26.7 percent) never attended any school and 0.8 million (12.1 percent) dropped out of schools in the last decade from 2001 to 2011[12].

Human Rights Watch, 2012 report mentions that a significant number of children with disabilities do not go to formal schools. They are mostly rejected during school admission and their parents too are unaware that education is a fundamental right for their children. Due to hurdles and challenges at school and in the home, children with disabilities have a high dropout rate [13]. 'While Nepal has made significant progress toward achieving universal primary education as part of its commitment to the Millennium Development Goals (MDGs), children from marginalized communities, such as children with disabilities, represent a significant portion of the 330,000 primary school aged children who remain out of school in Nepal,' according to the Nepalese government and the United Nations[14].

The 'Inclusive Education Policy for Persons with Disabilities, 2016' is currently in effect in Nepal. The policy states that students should be able to study in their native communities without prejudice, but it also allows for special education for children with disabilities [15]. However, the policy itself indicates that there is the problem in mainstreaming disabled children in education due to the low level of responsibility taken by family members, community and schools. There is the problem in attaining expected achievements in ensuring the quality of lives and independent livelihoods by the disabled children even if there is a social inclusion policy. Due to the ineffectiveness of peer learning and child-centric activities, there is a chance for social exclusion and education derailment. There is no systematic modern information technology to facilitate the learning process for impaired youngsters, which has hampered their learning [15].

According to a report by the RCRD and Save the Children (2014), Nepalese disabled children are denied access to education, basic health care, early intervention, rehabilitation, and a variety of other specific supports that they are entitled to under the law. They frequently confront infrastructural obstacles, societal discrimination and discriminatory ill-treatment in the home, and school rejection [16].

Through this literature, it can be figured out that there is definitely learning difficulty to the children with disabilities. However, what other literatures' findings suggest and validate is that the fact that there is learning difficulty among CWDs seems to be unknown. Thus, this paper intended to go through the literature based on the education of disabled children and analyze the situation of learning difficulties to the children in the entire world. The purpose of this exploratory method was to obtain qualitative results of the literature based on the education of children thereby unveil the empirical findings on learning difficulties faced by the CWDs.

In the above context, a research project to answer the research questions was developed as to what are the empirical evidence to validate that there are learning difficulties among the children with disabilities in schools and how are the literatures' findings shaped in debunking the problems and challenges faced by the children with disabilities?

Methods and Processes

This study is based basically on descriptive analysis in the findings of different studies carried out in the world from the perspectives of CWDs and the education of CWDs. The purposive idea of the study was to figure out the learning difficulties to the CWDs. Thus, purposively, the lens of the paper was inclined to learning difficulties to CWDs. There were three bases (Disability and education, children with disabilities and education, and inclusive education of CWDs) for the collection of the studies. The major purpose of determining the categories was to ensure the maximum no. of studies in the field of disabled-focused education.

With the view to find the answers to the above research questions, first relevant literatures regarding the findings of CWDs, which basically focused on their learning and educational difficulties to the children were reviewed. I went through different literatures written on CWDs. For that, I visited the websites of scholarly journal articles. I downloaded more than 60 studies mostly journals based on the thematic area then I skimmed the studies on its inclination to children with learning difficulties. Thus, this paper is the analysis of 52 studies from educators, learners and institutions perspectives.

After compiling literatures on the thematic areas, I summed up the key findings of the literatures. There were some challenges to collect disability based studies via search engines. There was availability of learning difficulties based studies in the search engines but the challenge was to get free access to most of the studies. So in that case, I contacted my international friends and asked them to download the journals.

The major findings of each study were also captured from the documents to explain the research findings in the area of CWDs and their learning difficulties in schools. I analyzed the findings by exploring the perspective of disabilities and thereafter tried to set the lens on findings through theoretical approach of learning difficulties among the CWDs.

Findings

A study on inclusive education in the Tanzanian context with a focus on Head Teachers' and Teachers' Perspectives suggested that the inclusive schools were with barriers that hindered effective implementation of inclusive education. The major obstructions included an inaccessible physical infrastructure, a similar curriculum, untrained teachers, and a lack of teaching and learning materials. It was found that majority of teachers did not support inclusive education because untrained teachers implement it [17].

In this line, another study focused on teachers' perspectives carried out in Canada elaborated four features of inclusive education from the perspectives of teachers (1) attitudes towards inclusion (2) supportive communication and collaboration (3) classroom community; and (4) support and training. The results of this study also corroborate the above results and indicate some differences between elementary and secondary teachers' understanding and perceptions. The secondary teachers have to some extent good understanding of inclusion and inclusive education [18].

The above findings are based on the barriers to the students because of the untrained teachers to implement the ideas of inclusive education. Similarly, there are infrastructural including material barriers that are curtailing the learning needs of the children. The barriers to the children are not only based to the schools but also to the teachers' self-efficacy.

It can be assumed that teachers engaged in teaching disabled children could be facing problems in doing so owing to diverse impairments. Teachers' sense of efficiency decreases as difficult students grow older. Here the disabled children are referred to as difficult children. The study further reveals that teachers do not reject hard-to-reach students; rather they think they are not teaching properly[19].

There are always the roles of teachers' self-efficacy, knowledge and attitude to provide better education in inclusive education settings. Teachers' self-efficacy is a crucial factor that drives students' motivation and explains their actions. Educators and academics have spent a lot of work attempting to figure out how to evaluate and comprehend the efficacy of instructors. In addition to describing teachers' behavior, researchers use self-efficacy as one of the factors used to predict motivation. [20]. There is another argument that self-efficacy is directly related with the learning achievements of the students. Baron and Byrne (2004) indicated that self -efficacy has an influence significant to the activity of learning. In the activity of learning, self-efficacy is associated with the belief the student will be its ability to perform tasks, organize activities to learn their own, and live with the hope of academics of their own and others [21]. Thus, self-efficacy is highly essential to achieve the successful task and the duties of the school.

Not only self-efficacy, the attitude and knowledge of teachers also play a role to ensure learning opportunities for the CWDs. Several researchers' focus on teachers' attitudes toward the inclusion of students with disabilities. It has been concluded that eventually, the teacher's attitude toward inclusion affects the learning environment of the student in the schools [22]. As concluded by Hellmich et al. (2019), attitudes, knowledge and self-efficacy are crucial in implementing high-quality inclusive education practices in the schools; we can say easily that there is role of self-efficacy, knowledge and attitude to implement inclusive education in the schools. The roles have been supported by other literatures either [23].

Teachers' attitudes towards different categories of disabilities may differ and this assumption was proved by the study on "Teachers of the deaf as compared with other groups of teachers: attitude towards people with disabilities and inclusion". It has revealed that the attitudes of teachers varied depending on their position and situation. Teachers of the deaf had a more favorable attitude than the other groups of teachers toward people with disabilities, but their attitude toward integration was the most negative [24].

In the case of hearing impairment on the learning process, it has been found that the hearing impaired students will have lower learning process due to their hearing impairment. As suggested by Manchaiah and Stephens (2011), hearing impairment causes a variety of psychosocial, mental, and physical effects that lead to their limitations on activity and restrictions on participation [25]. Powell, Hyde, and Punch (2014) have indicated that the hearing impaired students' academic participation is badly hampered by communication barriers [26].

It has been found that the students with hearing impairment in inclusive classes seem less responding to questioning, opinion making and involving themselves in class discussions. As to Stinson and Liu (1999), hearing impaired students need more time to understand the questions asked by their teachers, colleagues and to answer these questions correctly [27]. They need more time for group communication and interaction. Kyle (2006) says that students with hearing impairment will have more difficulty in following class discussions [28]. Thus, the teachers have a pivotal role in making such children understand and motivate them to get involved in class discussions. It solely depends on the way of teaching the students. What materials the teachers use in teaching will have a major role in understanding and receiving the information by the students. According to Charema (2010), the attitude and willingness of teachers to accommodate and attend to the needs of the students, thus cannot be underestimated [29]. In the learning process, students with hearing impairment are very much dependent on what is said by the teacher [30].

If this is the situation then what could be the recommendation? For recommendation as suggested by Talmor and Kayam (2011), mentioned that a single strategy was not sufficient in changing the attitudes of teachers and teacher trainers. The two strategies (instilling knowledge on disorders and exposure to individuals with special needs must be included [31].

Research on inclusive education policy, the general allocation model, and dilemmas of practice in primary schools has come up with the findings that inclusive education has not resulted in positive outcomes for students who need learning support. The research reveals that such a situation is because the built system on defective assumptions focuses on a psycho-medical perspective of disability where intersectionality of disability with class or culture is not considered. The study opines that those students who need support are better understood as 'home/school discontinuity' rather than disability. Further, the study uncovers the power of some parents to use social and cultural capital to ensure eligibility to enhanced resources. The study has argued that a hierarchical system has managed in mainstream schools to support needs in inclusive settings as a result of funding models [32].

The above description is the qualitative finding of the study. Quantitatively such findings might be different in terms of perceptions of educators towards inclusive education.

A study carried out on “Educators’ Perceptions of Inclusive Education”. It has been found that the educators' gender, qualifications and experiences as educators have no relation to their perceptions of the successful implementation of inclusive education and educators' perceptions of an inclusive classroom [33].

Another study on "Teachers' Attitude towards Inclusion of Students with Intellectual Disability in Community Schools" was carried out in Nepal to find out the acceptance of inclusive education model by the teachers of students with intellectual disabilities. The hypotheses were set for teachers’ attitude, subjective norms and perceived behavior.

The study concluded that generally teachers feel higher social pressure to practice inclusive education for students with intellectual disability (SWID). It further elaborated that teachers are less positive about the notion of providing inclusive education of (SWID). Similarly, the intention of teacher to practice inclusiveness in classrooms for SWID increases with their knowledge about intellectual disability. Intention of teachers to practice inclusiveness in the classroom was associated with teachers’ expectation from people with intellectual disability more that their knowledge about intellectual disability [34].

Through the above findings of inclusive education, it can be argued that teachers’ attitude, beliefs, intention, training, etc. are the major determining factors to practice inclusive education in the classroom.

A study on "Schooling of Girls with Disability" has the findings that school culture, the available resources, facilities and services at school were not as adequate as needed fully for Girls with Disabilities (GWDs) which has resulted for those girls to acquire knowledge and skills to their full potentials [35]. Despite these obstructions of structural constraints, the girls were found diligent in comparison to other peers seems like struggling to ensure their better future to lead a better life enthusiastically as suggested by human agency’s transformative capacity theory of Giddens [35].

The study was focused on GWDs but this study is rounded up to both males and females disabled children who are studying at resource class schools and special schools. The available resources, facilities, services provided by teachers will be the study realm. So, this finding can be linked on the findings of this study too.

In case of children with multiple disabilities, different studies have indicated that there are difficulties in learning to the children. Children with multiple disabilities require special support in educational services and special arrangements of education placement and curriculum design but these things basically lack in most of the cases [36]. Avramidis and Norwich (2002) came to the conclusion that teachers are more ready to make an attempt to include kids with mild disabilities, but this is not the case for students with more severe or numerous needs [37]. There are cases of exclusion, educational hurdles, and a lack of opportunity for kids with multiple disabilities to utilize the present educational system. [24].

Through these findings, we can argue that there are severe forms of learning difficulties to the multiple disabled children because of the approach to the inclusive education principle.

Regmi (2017) mentioned on practice of inclusive education in schools critically that there is less effective inclusive pedagogy within the available policies on inclusive education specially designed for children with disabilities in Nepal. Pedagogical practices have lapsed owing to a number of issues, including inadequate teachers, a lack of inclusive practice in schools, a lack of community and school coordination, and limited financial resources. Social issues such as social ideas and values, a lack of resources, and ignorance, as well as teachers' unfavorable approaches and attitudes, all play a role in inefficient inclusive education. [38].

In line with the findings, Thapaliya (2018) indicated that there are contradictions in contents of policies in Nepal that are more inclined to the medical approach of disability. There is a contradiction on government and society perspectives on disability where the society believes that disability is because of the evil deeds of the particular persons which will automatically contribute to discrimination, stigmatization, segregation and eventually exclusion from the society itself [39]. The attitudes of teachers and parents are found negative towards disability which is affected by sociocultural ideology, barriers in texts and curriculum, and confusing policies. It revealed that there are different significant influencing factors such as teacher type, age, gender, education levels, coursework, and residence on determining teachers’ attitudes towards inclusive education [40].

Access to school is another issue for the children with disabilities. Research by Oosterlee (2012) mentioned that access to schools differs among disabled children: the availability and accessibility of schools for deaf children are sufficient, while the availability of schools for blind and intellectually disabled children is doubtful [41]. There is no school for children with numerous disabilities. Because the location of schools, parents' income, and existing (special) schools all have an impact on each other, accessibility, cost, and availability are all intertwined. Due to free education for children with disabilities, income as an independent factor does not play a crucial role in school attendance. Furthermore, parents' educational attainment and caste origins are not strongly associated to the children's attendance at school [41].

The hurdles of learning difficulties are because of the lack of coordination between stakeholders; government inefficiency in a time of political upheaval and a tendency for education stakeholders to frame inclusion of disabled children as primarily a socio-economic issue to the neglect of other confounding factors [42]. Similarly, there are other important factors for a good life including education of the children with disabilities including their families. Physical and emotional health, as well as degrees of empowerment and independence, are critical. [43].

A baseline study of children with disability by UK Aid (2019) revealed that disability has the casual consequence of social exclusion and if it intersects with other forms of marginalization and discrimination, becomes increasingly hampering on daily lives of persons/children with disabilities [44].

Regarding disability and ensuring education, the sensitive issue is basically in its execution. The governance part is also weak in the organizations that are working on the issue of disability. The existence of policies is largely being shadowed due to lack of responsibility, accountability and transparency [45].

Major factors of the learning difficulties are the attitude towards education of disabled children, lack of resources in the schools to manage the overload of inclusive education coupled with the alarming factors of poor, poverty and deteriorated health in the developing countries like Nepal [46]. The current problem is the education opportunity as it relates to the treatment of integrated education for the education needs of children with disabilities. In Nepal, the issue of inclusion seems to have been partially met as special needs education is to some extent availed [47].

Discussion

The studies clearly inclined to learning difficulties to the children. These studies indicate that there is an obstruction in implementing the phenomena of inclusive education due to different factors associated with the parents, teachers, schools and even the community. A study conducted in USA revealed that there is a need of more effort from the teachers, peers and the schools to help students in the school environment to continue the education of disabled students [48].

Even though the studies inclination towards learning difficulties to the children, we cannot directly assume that there are not any learning opportunities for the CWDs. Due to learning opportunities through inclusive education and special education concepts, there are some positive changes to the education of the children. Due to inclusive education and special education practices, the learning difficulties are gradually changing to learning opportunities in some of the countries. It is established that the benefits of a specific interactive learning environment in terms of achieving the highest levels of school achievement and group cohesion for all students, as well as maximizing on the benefits of interaction for learning. [49; 50]. However, disability is more or less equated with learning difficulty in most of the countries because of the hindering factors associated with it. People with disabilities are subjected to various deprivations, according to the World Bank Report (2009), and they are the most excluded from school. It was also discovered that the more severe a child's condition, the less likely the child is to attend school. [51].

The Tanzanian study by Tungaraza (2014) indicates that there are hindering factors associated from physical barrier to the teaching learning materials in the school [17]. This finding is supported by a report of Canada that there are key learning barriers to the CWDs at the primary and secondary, or at the post-secondary level. Inadequate money, physical inaccessibility, time-consuming and inefficient accommodation processes, negative attitudes and stereotypes, and a lack of understanding of all parties' rights and obligations are the main roadblocks [52].

When we talk about another factor as teachers, in most of the schools basically in developing countries, there is lack of trained teachers on the concept and idea of inclusive education and other facilities and materials available in the school for the CWDs.

This indication was supported by a study in USA, which revealed that having a positive attitude toward inclusion can be challenging when teachers do not have basic skills (e.g., ability to modify the curriculum, understanding of student disabilities, managing challenging behaviors) necessary to facilitate inclusion [53].

Teachers' attitudes and perceptions of students with impairments are always important. If their perceptions and attitudes are positive, inclusive education or the education of CWDs will be positive and effective. In the education of children with disabilities, the teacher's attitude is critical since their judgments can have a social, emotional, and intellectual impact on a child's well-being. [54]. The attitudes of teachers according to their status and levels may differ as revealed by Murray et al. (2008). University faculty generally had positive perceptions about students with learning difficulties and was willing to spend time supporting students with learning difficulties [55].

In teachers’ self-efficacy, knowledge and attitude, several researches have revealed that there is relationship between teachers’ self-efficacy and inclusive education practices. A positive relationship between teachers’ attitude towards inclusion and their self-efficacy in practicing inclusion was reported in Tanzania [56]. Similarly, another study conducted in Canada indicated that higher self-efficacy for collaboration was the only predictor associated with more positive attitudes about inclusive education practices for students with developmental disabilities [57].

A study by Wang et al. (2012) in Shanghai, China reported that general and special education teachers differing in their self-efficacy for inclusion [58]. Teachers in the mainstream school reported lower efficacy for inclusive instructional strategies and collaboration, which was justified by the earlier observation (cited in 58) that minimal knowledge of teachers in general schools for catering to the diverse needs of children with disabilities as the biggest barrier to successful implementation of inclusive practices [58]. Wang et al. (2012) raised their concerns about the lack of training that general education teachers receive (both theoretical and practical) through their teacher education programs [58].

As the studies indicated that the teachers’ self-efficacies are not up to the level in most of the developing countries reason why there are learning difficulties to the children.

In quantitative approach, the finding suggests that for successful implementation of inclusive education for disabled children, gender, qualifications and experience of educators will not obstruct anything. These are not determining factors to run successfully the inclusive education in the schools.

The pedagogical approach also mattes to enhance the learning capabilities of disabled children. Further, the contractions in policy and societal perspectives to gauge the disability matters a lot to support learning efficiency of the children. Similarly, access to the school is another area whether to ensure learning opportunities to the disabled children. The better coordination between the stakeholders always contributes to ensuring learning opportunities for the children.

Teachers' pedagogical practices are a fundamental social justice issue in regard to improved learning outcomes for all children [59; 60]. Pedagogy is complicated and includes relationships between teachers, children, curriculum content, and knowledge created [61]. This relational perspective of pedagogy recognizes the importance of teacher–child relationships and relationships between children for effective pedagogical practices [62].

Lewis and Norwich (2005) explored about the complex relationship between teachers' knowledge, the curriculum and pedagogical strategies. They suggest that teacher education should include the study of child development and the psychology of learning and promote a holistic approach [63]. This view led to an increasing focus on 'inclusive pedagogy' in a range of countries [64; 65]. Thus, with regard to teaching strategies, Lewis and Norwich (2005) concluded that impairment-specific pedagogy was advisable [63]. They argued that the majority of students' needs are tried to meet through the adaptation of general teaching strategies catering for differences through 'degrees of deliberateness and intensity of teaching', which are not suitable for the CWDs [63].

The learning difficulties are there because of physical and mental health situation of the disabled children. The casual consequences of social exclusion contribute to deteriorated mental and physical health and eventually the learning difficulties of disabled children. The policy existence and its proper implementation are crucial to avail the rights of CWDs, which must be supported by the attitudes and beliefs of the educators and these are generally lacking in developing countries.

The social exclusion for the learning of the children is created by parents, teachers and the schools itself. Parents' mindsets, attitudes, and beliefs play a crucial role in their decision-making about whether or not to include or exclude their kid with a handicap, as well as in influencing policymakers and practitioners [66]. Scholars have also argued that there is a lack of understanding about disability and that professionals' instructional skills are lacking [67]. Similar incidents revealed a lack of dedication and preparation on the part of school staff to adjust teaching and learning materials. Furthermore, there is evidence that unqualified workers and facilitators contributed to the exclusion of disabled children [68]. As a result, parents prefer to care for their children at home, which obstructs the children's learning chances [68].

Another source of social isolation is gender stereotypes. In every culture, gender stereotypes combine with disability stereotypes to form a deep matrix of gendered disability, created within specific historical settings and changing those situations over time. Girls with disabilities are at the crossroads of many forms of disability and gender discrimination [54]. This situation ultimately limits the learning options for disabled girls.

Through all these findings, we can figure out that the learning difficulties among the disabled children are at an alarming stage and it is determined by different factors. A host of contributing factors has been identified for the curtailment of the learning opportunity of the children. Through the evidence of above literatures, it can be figured out that the determining factors are the roles and responsibilities of educational authority, availability of important knowledge on disabilities at schools and with the teachers. The other factors include provisions for the rights for the children, learning environment, approach on equality and inclusiveness in the schools including teachers' knowledge and attitude, their perceptions toward the students and their self-skills/efficacy to deal with the complexity of such children's learning needs including the role of parents. The attitude of parents toward disabled children and their education can be a major facilitator or a major impediment to inclusion and engagement in mainstream society, including schooling [54].

These factors were also spelled out by the inclusive education and educational theory of Knight, 1999. The theory has pointed out that there is a need for democratic authority, inclusiveness and democratic classroom, the democratic curriculum, student rights, the nature of participation in decisions that affect one's life, establishing optimum enabling environment for learning, and equality for the disabled children in the schools [69]. These requirements must be fulfilled to ensure quality education for children with disabilities. If these requirements are not met, children with disabilities are bound to face learning difficulties. The findings of the literatures clearly revealed that the above factors that ensure quality education for disabled children are glaringly missing.

The learning difficulties faced by the children with disabilities are because of the perspectives of society, community and individuals. In disability, Rioux (1997) mentioned that there are three perspectives;

Disability is viewed as a medical or physical problem that can be prevented or decreased through biological, medicinal, or genetic therapies under the biomedical approach.

Disability is viewed as an individual condition with an emphasis on how to treat the functional impairment it causes under the functional approach.

Disability as a consequence of how society is organized, and the relationship between society and the individual under the rights-outcome approach [70].

A rights-based approach to education demands a systematic effort to identifying and removing the barriers and blockages that obstruct access, as well as a rigorous method to demonstrating entitlement of every child to education. A commitment to inclusive education would embrace this dimensional approach so that the concept of inclusive education came and applied in most of the countries. It requires an understanding of inclusion as an approach to education for all children. This approach needs to be supported by a broad strategic commitment across government to create the necessary environment for ensuring the rights of CWDs, then only the right to education of the children can be ensured[4].

Actually, everyone's perspective should be right outcome or right based approach to enhance learning opportunities but different entities (community, society, individuals) see disabilities through different lenses so the problems associated with disabilities have failed to see any solutions and remain where they are.

These perspective discrepancies can be linked to Giddens' Structuration theory. As argued by Giddens (1984), an individual's autonomy is influenced by the structure of the society. Giddens (1984) argues that both, 'structure' and 'agency' are associated with 'society' and the 'individual' (p. 162). Giddens' theory seeks to show that the knowledgeable actions of human agents discursively and recursively form a set of rules and, practices and routines [71]. So, we have sensed through different studies that how CWDs are influenced or affected by the school structures in the set of rules and practices applied mostly in developing countries. Here, we can link the Giddens explanation of the interaction of human actors and social structures in providing or curtailing the learning opportunities to the children. Thus, we can say that how is the structure formed and how is it functioning by the actions and interactions of human will determine the learning opportunities for the children.

When we talk about learning opportunities for the children basically in the developing countries, the structural problems created by the human actors/agencies and social structures are there and these are evident from the literatures also, so there is learning difficulty related problems for the children with disabilities in schools.

Conclusion

Through the above findings, it can be argued that teachers' attitude, beliefs, intention, training, etc. are the major determining factors to practice inclusive education for the children with disabilities to enrich learning opportunities for the children. It is revealed that because of the authority, structure, school environment and behaviors of teachers, parents and the society as a whole, the learning environment for the children are being problematic. Similarly, a blanket approach or one size fits all approach while dealing all with types of CWDs is a problem that impedes the process of providing better access and ensuring learning capabilities for the children. In contrary, the available resources, facilities, services provided by teachers are the major factors to ensure the quality education of CWDs in the schools.

We can conclude based on the theoretical perspectives that learning spaces having structural problems influenced by the human actors and the structures of the schools itself lead to the obstruction of the learning processes and abilities of CWDs. It is recommended that the basic themes to ensure learning needs of CWDs as suggested by Knight (1999) in the schools can be fulfilled through joint efforts, cooperation and coordination among all stakeholders along with better structural adjustments in the schools basically of developing countries. The principle of collectivism (collective efforts of all stakeholders) and empowerment (SMC/PTA, teachers and parents) is the best way forward to ensure the rights to education of CWDs as prescribed by UNCRPD, 2006 and legal provisions of all countries.

This paper brought forward the major determining factors or loopholes to curtail learning opportunities of the CWDs through empirical evidences and theoretical lenses. Further elaboration on the figured out determining factors and its influence to ensure or curtail learning opportunities to the children can be conducted through participatory or emancipatory research approach of disability. Such research would help us to better understand the circumstances of learning difficulties experienced by the learners themselves.

References

[1] M. Subedi. Challenges to measure and compare disability: A methodological concern. Dhaulagiri Journal of Sociology and Anthropology, Vol. 30, NO. (3/4), 2012.

[2] E. Shrestha & A. Nilsson. Mainstreaming disability in the new development paradigm evaluation of Norwegian support to promote the rights of persons with disabilities. The Nepal country report, 2012.

[3] UNDP. Facts sheet on persons with disabilities, 2021 https://www.un.org/development/desa/disabilities/resources/factsheet-on persons-with-disabilities.html

[4] UNICEF. The right of children with disabilities to education: A rights-based approach to Inclusive Education, Position Paper. UNICEF Regional Office for CEECIS, Geneva , 2012.

[5] Central Bureau of Statistics (CBS). Nepal living standards survey 2010/11. Government of Nepal, 2011.

[6] Central Bureau of Statistics (CBS). National population census report 2011. Government of Nepal, 2012.

[7] Disability Resource Center. Disability atlas of Nepal. School of Arts, Kathmandu University, 2016.

[8] D. Pop. Education policy and equal education opportunities. Open Society Foundations, 2012.

[9] UNICEF. Press release, 2019. https://www.unicef.org/press-releases/75-cent-children-disabilities-eastern-and-central-europe-and-central-asia-left-out

[10] UNESCO. Education transforms lives, 2020. https://en.unesco.org/themes/education/

[11] S. Muputisi. Educators 'perceptions of the inclusion of learners with hearing impairment into advanced level classes at selected high schools in Gweru urban [Master's thesis]. Midlands State University, Zimbabwe, 2014.

[12] D. Singh. Dilemma and challenges of early education inclusion in schools of Lucknow, Uttar Pradesh, India. Asian Journal of Inclusive Education, Vol. 4, No. 1, pp. 51-77, 2016.

[13] Human Rights Watch. Education for disabled in Nepal, 2012 http://southasia. oneworld.net/resources/education-for-disabled-innepal#.U0pYnlWSzL8

[14] K. Chalise. Perceptions of teachers toward inclusive education focus on hearing impairment. World Academy of Science, Engineering and Technology International Journal of Educational and Pedagogical Sciences, Vol. 14, No. 1, 2020.

[15] Department of Education (DoE). Inclusive education policy for persons with disabilities. Author, 2016.

[16] Resource Center for Rehabilitation and Development, & Save the Children. Disability service mapping with special focus to children with disabilities. Authors, 2014.

[17] F. D. Tungaraza. The arduous march toward inclusive education in Tanzania: Head teachers' and teachers' perspectives. Africa Today, Vol. 61, No. 2, pp. 109-123, 2014.

[18] D. Richmond, A. Irvine, T. Loreman, J. Lea Cizman & J. Lupart . Teacher perspectives on inclusive education in rural Alberta, Canada. Canadian Journal of Education, Vol. 36, No. 1, pp. 195-239, 2013.

[19] J.A. Lopes, I. Monteiro, V. Sil, R.B. Rutherford & M.M. Quinn. Teachers' perceptions about teaching problem students in regular classrooms. Education and Treatment of Children, pp. 394-419, 2004.

[20] R. M. Klassen & M. M. Chiu. The occupational commitment and intention to quit of practicing and pre-service teachers: Influence of self-efficacy, job stress, and teaching context. Contemporary Educational Psychology, Vol. 36, pp. 114–129, 2011.

[21] R. Baron, S. Byrne. Social psychology. Jakarta: Erlangga, 2004.

[22] A. K. Van Reusen, A . R. Shoho & K. S. Barker. High school teacher attitudes toward inclusion. The High School Journal, pp. 7-20, 2000.

[23] F. Hellmich, M. F. Loper & G. Gorel. The role of primary school teachers' attitudes and self-efficacy beliefs for everyday practices in inclusive classrooms- A study on a verification of 'the planned behavior theory'. Journal of Research in Special Educational Needs, Vol. 19, No. 1, pp. 36–48, 2019.

[24] K. Lampropoulou. The education of multiple disabled children and adults in Greece: The voices and experiences of parents and parent associations [Doctoral dissertation]. University of Birmingham, Greece, 2012.

[25] V.K.C. Manchaiah & D. Stephens. Models to represent communication partners within the social networks of people with hearing impairment. Audiological Medicine, Vol. 9, No. 3, pp. 103–109, 2011. http://www.tandfonline

[26] D. Powell, M. Hyde & R. Punch. Inclusion in postsecondary institutions with small numbers of deaf and hard-of-hearing students: Highlights and challenges. Journal of Deaf Studies and Deaf Education, Vol. 19, No. 1, pp. 126-140, 2014.

[27] M. Stinson & Y. Liu. Participation of deaf and hard-of-hearing students in classes with hearing students. Journal of Deaf Studies and Deaf Education, Vol. 4, No. 3, pp. 191-202, 1999. https://academic.oup.com/jdsde/article/4/3/191/440895

[28] J. G. Kyle. Integration of deaf children. European Journal of Special Needs Education, Vol. 8, No. 3, pp. 201-220, 2006. http://www.tandfonline.com/doi/abs/10.1080/0885625930080303

[29] J. Charema. Inclusion of primary school children with hearing impairments in Zimbabwe. Africa Education Review, Vol. 7, No. 1, 2010. http://www.tandfonline.com/doi/abs/10.1080/18146627.2010.48581 0

[30] J. D. Smith. Sekolah inklusif: Konsep dan penerapan pembelajaran (Denis, Ny. Enrica, Trans). Nuansa Cendekia (Original work published 1998), 2012.

[31] R. Talmor & O. Kayam. Changing teachers' attitudes towards students with special needs and their inclusion: Research findings. Issues in Special Education & Inclusion, pp. 34-21, 2011.

[32] E. Margaret. Inclusive education policy, the general allocation model and dilemmas of practice in primary schools [Doctoral dissertation]. National University of Ireland, 2013.

[33] S. Siebalak. Educators' perceptions of inclusive education [Doctoral dissertation]. University of Zululand, South Africa, 2002.

[34] S. Shrestha. Teachers' attitude towards inclusion of students with intellectual disability in community schools [MPhil. dissertation]. Kathmandu University School of Education, Nepal, 2017.

[35] B.S. Thapa. Schooling of girls with disability: A phenomenological study of Nepali girls [Doctoral dissertation]. Kathmandu University, Nepal, 2012.

[36] Y. Chen. Education to multiple disabilities in China. The 24th Asia-Pacific International Seminar on Special Education, Yokosuka, Japan, 2004.

[37] E. Avramidis & B. Norwich. Teachers' attitudes towards integration/inclusion: A review of the literature. European Journal of Special Needs Education, Vol. 17, No. 2, 2002.

[38] N. P. Regmi. Inclusive education in Nepal from theory to practice [Doctoral dissertation]. Ludwig-Maximilians-University, Munich, 2017.

[39] MP. Thapaliya. Moving towards inclusive education: How inclusive education is understood, experienced and enacted in Nepali higher secondary schools [Doctoral dissertation]. University of Canterbury, New Zealand, 2018.

[40] S. Aryal. Teachers' attitude towards inclusive education in Nepal [Doctoral dissertation]. Graduate School of Daegu University, Korea, 2013.

[41] AS. Oosterlee. The accessibility to basic education for disabled children in Baglung district, Nepal [Master thesis]. Utrecht University, the Netherlands, 2012.

[42] S. John. Inclusion for all? An analysis of inclusive education strategies for marginalized groups in Nepal [Master thesis]. Graduate School of Social Sciences University of Amsterdam, 2018.

[43] L. Brandt. The inclusiveness of society for children with disability in Nepal. Master Thesis, Maastircht University, Faculty of Health, Medicine and Life Sciences, Netherlands. Karuna Foundation, 2015.

[44] UK AID. Of children with disability. International Disability Alliance, 2019.

[45] R. Baral. Historical policy review on disability. Research Nepal Journal of Development Studies, Vol. 1, No. 1, pp. 73-82, 2018.

[46] L. M. Banks, M. Zuurmond, A. Monteath–Van Dok, J. Gallinetti & N. Singal. Perspectives of children with disabilities and their guardians on factors affecting inclusion in education in rural Nepal: "I feel sad that I can't go to school". Oxford Development Studies, Vol. 47, No. 3, pp. 289-303, 2019.

[47] M. Sugimura & M. Takeuchi. Rethinking implications of inclusive and special needs education in the context of Nepal [USJI Seminar]. Sophia University, JICA Research Institute, 2017.

[48] R. Wieringo. A case study of the experiences of students with disabilities who did not complete high school [Doctoral dissertation]. Liberty University, Lynchburg, VA, USA, 2015.

[49] R. Valls & L. Kyriakides. The power of interactive groups: how diversity of adults volunteering in classroom groups can promote inclusion and success for children of vulnerable minority ethnic populations. Cambridge J. Educ. Vol. 43, pp. 17–33, 2013.

[50] A. Aubert, S. Molina, T. Shubert & A. Vidu. Learning and inclusivity via interactive groups in early childhood education and care in the Hope school. Spain. Learn. Cult. Soc. Interact, Vol. 13, 90–103, 2017.

[51] The World Bank. (2009). People with Disabilities in India: from commitments to Outcomes. New Delhi: Human Development Unit, South Asia Region.

[52] Ontario Human Rights Commission. (2002). Policy and guidelines on disability and the duty to accommodate. www.ohrc.on.ca.

[53] R.A. Allday, S. Neilsen-Gatti & T.M. Hudson. Preparation for inclusion in teacher education pre-service curricula. The Journal of the Teacher Education Division of the Council for Exceptional Children, Vol 36, No. 4, pp. 298-31, 2013.

[54] S. Limaye. Factors influencing the accessibility of education for children with disabilities in India. Global Education Review, Vol. 3, No. 3, pp. 43-56, 2016.

[55] C. Murray, C. T. Wren & C. Keys. University faculty perceptions of students with learning disabilities: Correlates and group differences. Learning Disability Quarterly, Vol 31, No. 3, pp. 95-113, 2008.

[56] R.H. Hofman, J.S. Kilimo. Teachers' attitudes and self-efficacy towards inclusion of pupils with disabilities in Tanzanian schools. Journal of Education and Training, Vol. 1, No. 2, pp. 177, 2014.

[57] A. Montgomery, P. Mirenda. Teachers' self-efficacy, sentiments, attitudes, and concerns about the inclusion of students with developmental disabilities. Exceptionality Education International, Vol. 24, No. 1, pp. 18-32, 2014.

[58] M. Wang, F. Zan, J. Liu, C. Liu & U. Sharma. A survey study of Chinese in-service teachers' self-efficacy about inclusive education. Journal of International Special Needs Education, Vol. 15, No. 2, pp. 109 – 116, 2012.

[59] R. Kershner. Learning in inclusive classrooms. In P. Hick, R. Kershner, & P. Farrell (Eds.), Psychology for inclusive education: New directions in theory and practice, pp. 52-65. Abingdon, UK: Routledge, 2009.

[60] T. Majoko. Regular teacher preparation for inclusion. International Journal of Special Education, Vol. 32, pp. 207-236, 2017.

[61] R. Tinning. Pedagogy and human movement: Theory, practice, research. London, England: Routledge, 2010.

[62] J. T. Lysaker & S. Furuness. Space for transformation: Relational, dialogic pedagogy. Journal of Transformative Education, Vol. 9, pp. 183-197, 2011.

[63] A. Lewis & B. Norwich. Special teaching for special children: Pedagogies for inclusion. Maidenhead: Open University Press, 2005.

[64] L. Florian & K. Black-Hawkins. Exploring inclusive pedagogy. British Educational Research Journal, Vol. 37, No. 5, pp. 813–828, 2011.

[65] L. Florian & H. Linklater. Preparing teachers for inclusive education: Using inclusive pedagogy to enhance teaching and learning for all. Cambridge Journal of Education, Vol. 40, No. 4, pp. 369–386, 2010.

[66] M. Tryfon, A. Anastasia & R. Eleni. Parental perspectives on inclusive education for children with intellectual disabilities in Greece. International Journal of Developmental Disabilities, 2019.

[67] J. Siska, Y. Bekele, J. Beadle-Brown & J. Zahorik. (2020). Role of resource centers in facilitating inclusive education: Experience from Ethiopia. Disability & Society, Vol. 35, No. 5, pp. 811–830, 2020.

[68] L. Stevens & G. Wurf. Perceptions of inclusive education: A mixed methods investigation of parental attitudes in three Australian primary schools. International Journal of Inclusive Education, 24(4), 351–365, 2020.

[69] T. Knight. Inclusive education and educational theory, inclusive for what? [Paper presentation]. Paper presented at the British Educational Research Association Conference, University of Sussex, at Brighton, 1999. http://www.leeds.ac.uk/educol/documents/000001106.htm

[70] M. H. Rioux. When myths masquerade as science. In L. Barton & M. Oliver (eds.), Disability studies: Past, present and future. The Disability Press, 1997.

[71] A. Giddens. The theory of structuration, the constitution of society: Outline of the theory of structuration. Oxford Polity Press, 1984. https://www.journals.uchicago.edu/doi/abs/10.1086/228358?journalCode=ajs

Application of Electrohydrodynamic Printing for Preparation of a Flexible Magnetic Structure

M. Khandaei, M. Rafienie, S. Bonakdar, L.Hao, S.A. Poursamar

***Abstract*—**The present work, reports on combination of precise positioning ability of 3D printers with near field electrospinning (NFES) as a low-cost and scalable approach for producing a flexible magnetic structure. Using this high speed and large-area printing technique we can overcome the drawbacks of conventional electrospinning such as fiber structural inhomogeneity, random orientation, and non-reproducible results. In addition, the application of more complicated and expensive methods such as lithography or e-beam lithography or nanoimprint lithography can be avoided using this technique. The most important advantage of the reported fabrication method relative to electro-spinning process is the ability to have a special control on the distribution pattern of the electrodes on the surface which can lead to better control over the ratio of conductivity and transparency of the surface.

***Keywords*—**3D printing, Near-field electrospinning, Fe_3O_4 nanoparticles, Electrohydrodynamic printing, Flexible

1. Introduction

Recently a new fabrication method is reported as a combination of additive manufacturing and electrospinning which can provide significantly higher order into the prepared structure (Wang et al., 2008). In many literatures, this method is referred to as mechano electrospinning, electro hydro dynamic printing (EDH) (Yin et al., 2018) or near-field electrospinning (NFES) (Min et al., 2013). NFES is a powerful technique which recently developed to print out uniaxially aligned fibers with precise control of fiber size and placement (Min et al., 2013; Fuh and Lu, 2014). The key advantage of this technique lies in the low applied voltage and short electrode-to-collector distance which provide better fiber deposition controllability.

Electrohydrodynamic printing provides higher resolution than standard printing techniques such as inkjet printing. The higher resolution is achieved by employing a different method to form the droplet: by applying an electric field between a nozzle (with typical inner diameter between ~100 nm to several μm) and the substrate.

In Electrohydrodynamic printing modes, the resulting ink droplets are in an electric field between the tip and the substrate and are essentially pulled out of the nozzle tip onto the substrate. The pulling of the ink droplet out of the nozzle in EHD printing, in contrast to the pushing of the ink droplets out of the nozzles in inkjet printing, makes EHD more flexible because the pulling strength in EHD can be easily tuned by altering the imposed electric field. EHD also enables an accurate droplet placement due to the assisting electric field lines. Note that, although inks with variety of different electrical conductivities have been shown to be EHD printable, the need for an electric field does somewhat limit the method in terms of substrate and the ink characteristics.

Near-Field Electrospining is another typical mode of electrohydrodynamic printing when the solution has a low viscosity. The jet breaks up into droplets with diameter ranging from a few nanometers to hundreds of micrometers. The challenge of optimizing the conductivity properties of such materials is network formation methods. A low junction resistance between nanostructures is important for decreasing the sheet resistance and the junction resistances between deposited electrodes can increase without coherent order within the network structure.

However, all these works state that the electrospinning is a random and non-controllable process, so that, the orientation and distribution of the polymer nanofiber template are difficult to control, and there is little to none control over fabrication process in electrospinning process which can results in higher electrical resistance and lower transparency and sample-to-sample variation of electrode performance may not be avoided.

The present work reports on combination of precise positioning ability of 3D printers with NFES as a low-cost and and acceptable feature for producing a flexible magnetic structure.

2. Materials and Methods

2.1. Materials

polycaprolactone with molecular weight of 80,000 g·mol^{-1} was purchased from Tabriz Petrochemical Company (TPCO, Tabriz, Iran). Adequate amount of PCL was dissolved in chloroform at 50 °C for 2 hour in order to yield 14% w/v solution. The viscosity of these solutions was measured at room temperature by using a digital rotational viscometer. In order to reach the final solution, 2.5%, 5% and 10% of Fe_3O_4 were added to the solution, respectively. Also prepare 0% solution as a control group.

2.2. NFES Printing of Microscale Fiber

As shown in Figure 1, a custom 3D electrohydrodynamic 3D printing platform based on the Choc Creator 1 platform (Choc Edge Ltd, UK) was used to build a well-aligned microfiber mask on flexible substrates. The PCL solution was placed in a 1 ml syringe. The print nozzle with an internal diameter of 600 μm was connected to the positive terminal of a DC voltage generator and the voltage was set to 1.5 volts. A collector on the moving stage of the collector was connected to the negative terminal of the DC voltage generator. The distance from the

Mansoureh Khandaei is with the Isfahan University of Medical Sciences, Iran (e-mail: mansooreh.kh1377@gmail.com).

nozzle to the slide on the collector was set to the minimum and the speed was 3000 mm/s.

2.3. *Preparation of Printed Structure*

The PCL and Fe_3O_4 fibers were printed on slide placed on the collector and produced a special dimensional pattern. After complete drying, the printed pattern is removed from the slide surface for electrospinning.

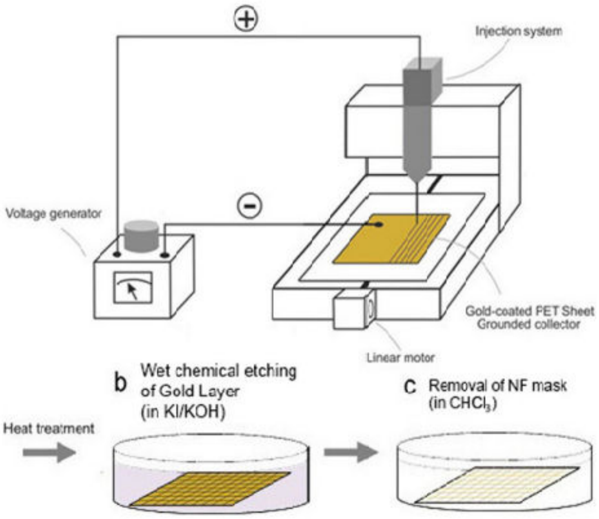


Fig. 1. The schematic view of NFES printing of microfibers.

2.4. *Electrospining*

For electrospinning on the surface of the printed pattern, 1 ml of solution with similar proportions of the solution prepared for printing is prepared and the distance between the needle tip and the collector is set at 20 cm. Finally, electrospinning is performed at a voltage of 20 kV and a rate of 0.01 ml/min at room temperature.

2.5. *Transparency Measurements*

The optical transparency at a wavelength between 200 and 1100 µm was measured using a UV–visible–near infrared spectrophotometer (SPECORD 250, Analytik Jena, Germany) in transmittance mode using 0.5 nm slit and the sample size of 25 × 25 mm.

2.6. *Bending Test*

A 20 × 20 mm sample was cut and was fixed between two holders. The Bending test was performed at 18 mm·s^{-1} rateand under the lateral dislocation of 11 mm causing a matching radius bending. The impact of 10,000 bending cycles on the conductivity of the samples was studied. The bending test was carried out on the same custom-made printer used for NFES by removing the extrusion section and installing clamps to hold the electrode specimen. The printer was programmed to oscillate 11 mm in Y-direction for up to 2500 cycles. The sheet resistance was measured using the method described in previous section and then test was continued for other 25,000 cycles for the total cycles of 10,000.

2.7. *Surface analysis*

Microstructure analyses were carried out using scanning electron microscope (SEMVEGA/TESCAN, Czech Republic). The samples were sputter coated (K450X Sputter Coater, EMITECH, UK) with a thin layer of gold (~60 nm, 3 fold sputter time 2 min at 20 mA) to improve the conductivity of the surface. The samples were studied under 20 kV acceleration voltage.

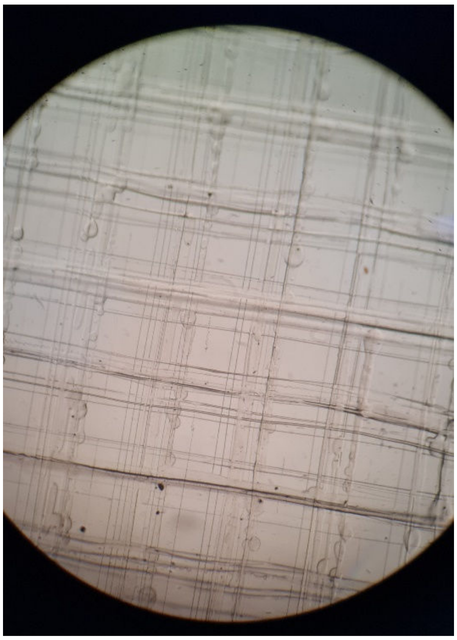


Fig. 2. Microscopic view from PCL – Fe_3O_4 printed structure

2.8. *TEM analysis*

TEM microscopes can produce images with a minimum magnification of 2,000 and a maximum magnification of 50 million and a resolution of 50 pm. Therefore, these types of microscopes are one of the best choices for studying the microstructure of various materials, especially different nanomaterials. In order to perform this analysis, samples with dimensions of 0.5 x 0.5 mm are placed on a grid and an electron beam is passed through the sample.

2.9. *Measurement of magnetic properties*

In order to measure and study the properties and magnetic behavior of the printed patterns, VSM analysis was performed on samples with dimensions of 0.5 x 0.5 mm. The applied magnetic field in the analysis is 0.8T and the results are presented in the form of a graph called the residual curve.

ACKNOWLEDGMENT

The authors wish to acknowledge the technical support of Choc Edge Ltd. in this research study. The printing technology used

in this study is based on Choc Creator 1 machine developed by Choc Edge Ltd (UK).

REFERENCES

[1] Askari, H., Fallah, H.R., Askari, M., Charkhchi Mohmmadieyh, M., 2014. arXiv:1409. 5293 [cond-mat.mtrl-sci].

[2] Azuma, K., Sakajiri, K., Matsumoto, H., Kang, S., Watanabe, J., Tokita, M., 2014. Mater. Lett. 115, 187–189.

[3] Fuh, Y.K., Lien, L.C., 2013. Nanotechnology 24, 55301–55308.

[4] Fuh, Y.-K., Lu, H.-Y., 2014. J. Micro/Nanolithography, MEMS, MOEMS 13, 043014.

[5] Gao, M.-Z., Job, R., Xue, D.-S., Fahrner, W.R., 2008. Chin. Phys. Lett. 25, 1380–1383. Hecht, D.S., Hu, L., Irvin, G., 2011. Adv. Math. 23, 1482–1513.

[6] Huh, J.W., Lee, D.K., Jeon, H.J., Ahn, C.W., 2016. Nanotechnology 27, 475302. Kang, M.G., Park, H.J., Ahn, S.H., Guo, L.G., 2010. Sol. Energy Mater. Sol. Cell 94, 1179–1184.

[7] Kim, W.K., Lee, S., Lee, D.H., Park, I.H., Bae, J.S., Lee, T.W., Kim, J.W., Park, J.H., Cho, Y.C., Cho, C.R., Jeong, S.Y., 2015. Sci. Rep. 5, 10715.

[8] Lee, J., Lee, P., Lee, H., Lee, D., Lee, S.S., Ko, H., 2012. Nanoscale 4, 6408–6414.

[9] Mazur, M., Kaczmarek, D., Domaradzki, J., Wojcieszak, D., Song, S., Placido, F. (Eds.), October 2010. Proceedings of 8th International Conference on Advanced Semiconductor Devices and Microsystems, ASDAM 2010. IEEE.

[10] Min, T.W., Kim, S.Y., Kim, T.S., Cho, B.J., Noh, H., Yang, Y.Y., Cho, H., Lee, J.H., 2013. Nat. Commun. 4, 1773–1781.

[11] Na, S.I., Kim, S.S., Jo, J., Kim, D.Y., 2008. Adv. Math. 20, 4061–4067.

[12] Nogi, M., Karakawa, M., Komoda, N., Yagyu, H., Nge, T., 2015. Sci. Rep. 5, 17254–17260.

[13] Otagawa, T., Madou, M.J., Wachsman, L.A., 1995, US Patent 5403680.

[14] Pasquier, A.D., Unalan, H.E., Kanwal, A., Miller, S., Chhowalla, M., 2005. Appli. Phys. Lett. 87, 203511.

[15] Wang, X., Zhi, L., Mullen, K., 2008. Nano Lett. 8, 323–327.

[16] Wu, H., Kong, D., Ruan, Z., Hsu, P.C., Wang, S., Yu, Z., Carney, T.J., Hu, L., Fan, S., Cui, Y., 2013. Nat. Nanotechnol. 8, 421–425.

[17] Yin, Z., Huang, Y., Duan, Y., Zhang, H., 2018. Electrohydrodynamic Direct-Writing for Flexible Electronic Manufacturing. Springer.

[18] Zhang, D., Ryu, K., Liu, X., Polikarpov, J., 2006. Nano Lett. 6, 1880–1886.

Atypical Intoxication Due to Fluoxetine Abuse with Symptoms of Amnesia

Bilen, Ayse Gul

Abstract**—** Selective serotonin reuptake inhibitors (SSRIs) are commonly prescribed antidepressants which are used clinically for treatment of anxiety disorders, obsessive-compulsive disorder (OCD), panic disorders and eating disorders. The first SSRI, fluoxetine (sold under the brand names Prozac and Sarafem among others), had an adverse effect profile better than any other available antidepressant when it was introduced because of its selectivity for serotonin receptors. They have been considered almost free of side effects and have become widely prescribed, however questions about the safety and tolerability of SSRIs have emerged with their continued use.

Most SSRI side effects are dose related and can be attributed to serotonergic effects such as nausea. Continuous use might trigger adverse effects such as hyponatremia, tremor, nausea, weight gain, sleep disturbance and sexual dysfunction. Moderate toxicity can be safely observed in the hospital for 24 hours, and mild cases can be safely discharged (if asymptomatic) from the emergency department once cleared by Psychiatry in cases of intentional overdose and after 6 to 8 hours of observation.

Although fluoxetine is relatively safe in terms of overdose, it might still be cardiotoxic and inhibit platelet secretion, aggregation, and plug formation. There have been reported clinical cases of seizures, cardiac conduction abnormalities, and even fatalities associated with fluoxetine ingestions. While the medical literature strongly suggests that most fluoxetine overdoses are benign, emergency physicians need to remain cognizant that intentional, high-dose fluoxetine ingestions may induce seizures and can even be fatal due to cardiac arrhythmia.

Our case is a 35-year old female patient who was sent to ER with symptoms of confusion, amnesia and loss of orientation for time and location after being found wandering in the streets unconsciously by police forces that informed 112. Upon laboratory examination, no pathological symptom was found except sinus tachycardia in the EKG and high levels of aspartate transaminase (AST) and alanine transaminase (ALT). Diffusion MRI and computed tomography (CT) of the brain all looked normal. Upon physical and sexual examination, no signs of abuse or trauma were found. Test results for narcotics, stimulants and alcohol were negative as well. There was presence of dysrhythmia which required admission to the intensive care unit (ICU).

The patient gained back her conscience after 24 hours. It was discovered from her story afterwards that she had been using fluoxetine due to post-traumatic stress disorder (PTSD) for 6 months and that she had attempted suicide after taking 3 boxes of fluoxetine due to loss of parent. She was then transferred to the psychiatric clinic.

Our study aims to highlight the need to consider toxicologic drug use, in particular the abuse of selective serotonin reuptake inhibitors (SSRIs), which have been widely prescribed due to presumed safety and tolerability, for diagnosis of patients applying to the emergency room (ER).

Keywords**—** Abuse, Amnesia, Fluoxetine, Intoxication, SSRI

Ayse Gul Bilen is with the Yildirim Bayezit University, Yenimahalle Education and Research Hospital, Child Abuse Center, Türkiye (e-mail: aysegulbilen@yahoo.com).

The Aesthetic Reconstruction of Post-Burn Eyebrow Alopecia with Bilateral Superficial Temporal Artery Island Scalp Flap

Kumar Y., Suman D., Sumathi

***Abstract*—** Introduction: Burns to the face account for between one-fourth and one-third of all burns. The loss of an eyebrow due to a burn or infection can have negative physical and psychological consequences for patients because eyebrows have a critical functional and aesthetic role on the face. Plastic surgeons face unique challenges in reconstructing eyebrows due to their complex anatomy and variations within genders. As a general rule, there are three techniques for reconstructing the eyebrow: superficial temporal artery island flap, a composite graft from the scalp, and mini or micro follicular grafts from the scalp. In situations where a sufficient amount of subcutaneous tissue is not available and the defect is big such as the case of burns, flaps like the superficial temporal artery scalp flap remain reliable options. In 2018, a 17-year-old female patient presented to the department of Burns Plastic and reconstructive Surgery of Guru Teg Bahadur Hospital, Delhi, India. A scald-burn injury to the face occurred two years before admission, resulting in bilateral eyebrow loss. We reconstructed the bilateral eyebrows using bilateral scalp island flaps based on the posterior branch of the superficial temporal artery. The reconstructed eyebrows successfully assumed a desirable shape and exhibited a natural appearance, which was consistent with preoperative expectations and the patient stated that she was more comfortable with her social relationships. Among the current treatment procedures, the superficial temporal artery island flap continues to be a versatile option for reconstructing the eyebrows after alopecia, especially in cases of burns. Results: During the 30 days follow-up period, the scalp island flap remained vascularised with normal hair growth, without complications. The reconstructed eyebrows successfully assumed a desirable shape and exhibited a natural appearance; the patient stated that she was more comfortable with her social relationships. Conclusion: In this case report, we demonstrated how scalp island flaps pedicled by the superficial temporal artery could be performed very safely and reliably to create new eyebrows.

***Keywords*—** alopecia, burns, eyebrow, flap, superficial temporal artery.

Yashvinder Kumar Kumar is with the University College of Medical Sciences, Delhi University, India (e-mail: dryashmamc@gmail.com).

Fresh Amnion Layer Grafting for the Regeneration of Skin in Full Thickness Burn in Newborn - Case Report

Priyanka Yadav, Umesh Bansal, Yashvinder Kumar

***Abstract*—** The placenta is an important structure which provides oxygen and nutrients to the growing fetus in utero. It is usually thrown away after the birth but it has a therapeutic role in regeneration of tissue. It is covered by the amniotic membrane which can be easily separated into the amnion layer and the chorion layer. The amnion layer act as a biofilm for the healing of burn wound and non healing ulcers. The freshly collected membrane has stem cells, cytokines, growth factors, and anti-inflammatory properties which act as a biofilm for the healing of wounds. It functions as a barrier and prevent heat and water loss and also protect from bacterial contamination thus supporting the healing process. The application of Amnion membranes have been successfully used for wound and reconstructive purposes since decades. It is very cheap and easy process and has shown superior results than allograft and xenograft. However there are very few case reports of amnion membrane grafting in newborn, we intend to highlight its therapeutic importance in burn injuries in newborn.

We present a case of 9 days old male neonate who presented to the neonatal unit of Maulana Azad Medical College with complain of fluid filled blisters and burn wound on body since 6 days. He was born at outside hospital at 38 weeks of gestation to a 24-year-old primigravida mother by vaginal delivery. The presentation was cephalic and the amniotic fluid was clear. His birth weight was 2800 gm and APGAR scores were 7 and 8 at 1 and 5 minutes respectively. His anthropometry was appropriate for gestational age. He developed respiratory distress after birth requiring oxygen support by nasal prongs for 3 days. On day of life 3 he developed blisters on his body starting from then face then over the back and perineal region.

At presentation on day of life 9 he had blisters and necrotic wound on right side of the face, back, right shoulder and genitalia affecting 60% of body surface area with full thickness loss of skin. He was started on intravenous antibiotics and fluid therapy. Pus culture grew Pseudomonas aeuroginosa for which culture specific antibiotics were started. Plastic surgery reference was taken and regular wound dressing was done with antiseptics. He had storming course during the hospital stay. On day of life 35 when baby was hemodynamically stable amnion membrane grafting was done on the wound site. For the grafting fresh amnion membrane was removed under sterile conditions from the placenta obtained by caesarean section. It was then transported to the plastic surgery unit in half an hour in a sterile fluid where the graft was applied over the infant's wound. The amnion membrane grafting was done twice in two weeks for covering the whole wound area. After successful uptake of amnion membrane, skin from the thigh region was autografted over the whole wound area by Meek technique in a single setting. The uptake of autograft was excellent and most of the areas were healed. In some areas there was patchy regeneration of skin so dressing was continued. The infant was discharged after three months of hospital stay and was later followed up in plastic surgery unit of the hospital.

***Keywords*—** Amnion membrane grafting, autograft, Meek technique, newborn.

Priyanka Yadav was with Department of Pediatrics, Maulana Azad Medical College, Delhi. She is now with Department of Neonatology, B. J Wadia Hopital, Mumbai, India (phone: +91-9315434237;e-mail: pysn2726@gmail.com).

Umesh Bansal, was with Department of Plastic Surgery , Maulana Azad Medical College, Delhi.. He is now with the Department of Plastic and Reconstructive Microvascular Suregry, Bhagwan Mahaveer Cancer and Research Center, Rajasthan, India. (e-mail:dr.umesh.bansal.bmchrc.com).

Yashviner Kumar, is with the Department of Burn and Platic Surgery, University College of Medical Science, Delhi, India.(e-mail: dryashmamc@gmail.com).

One Session Treatment (Ost) Is Equivalent to Multi-Session Cognitive Behavioural Therapy (Cbt) in Children with Specific Phobias (Aspect): Results for the UK, Non-inferiority, Randomised Controlled Trial with a Qualitative and a Health Economic Component

Barry Wright, Lucy Tindall, Alex Scott, Ellen Lee, Cindy Cooper, Katie Biggs, Penny Bee, Han-I Wang, Lina Gega, Emily Hayward, Kiera Solaiman, Dawn Teare, Thompson Davis, Jon Wilson, Karina Lovell, Dean McMillan, Amy Barr, Hannah Edwards, Jennifer Lomas, Chris Turtle, Steve Parrott, Catarina Teige, Tim Chater, Rebecca Hargate, Shezhad Ali, Sarah Parkinson, Simon Gilbody, David Marshall

***Abstract*—** Background: 5% to 10% children and young people (CYP) have specific phobias that impact upon daily functioning. Cognitive Behaviour Therapy (CBT) is recommended but has limitations. One Session Treatment (OST), a low-intensity alternative incorporating CBT principles, has demonstrated efficacy. Alleviating Specific Phobias Experienced by Children Trial (ASPECT) investigated the non-inferiority of OST compared to multi-session CBT for treating specific phobias in CYP. Methods: ASPECT was a pragmatic, multi-centre, non-inferiority randomised controlled trial in 26 CAMHS sites, three voluntary agency services and one university-based CYP well-being service. CYP aged 7- 16 years with specific phobia were randomised 1:1 to OST or CBT. Non-inferiority was assessed six-months post-randomisation using the Behavioural Avoidance Task (BAT). Secondary outcome measures included the Anxiety Disorder Interview Schedule, Child Anxiety Impact Scale, Revised Children's Anxiety Depression Scale, goal-based outcome measure, EQ-5DY and CHU-9D, collected blind at baseline and six months. An economic evaluation and qualitative study were undertaken. Results: 268 CYP were randomised to One Session Treatment (OST) (n=134) or CBT (n=134). Mean BAT scores at six-months were similar across groups in both intention-to-treat (ITT) and per-protocol (PP) populations (CBT: 7.1 (ITT, n=76), 7.4 (PP, n=57), OST: 7.4 (ITT, n=73), 7.6 (PP, n=56), on the standardised scale adjusted mean difference for CBT compared to OST -0.123, 95% CI -0.449 to 0.202 (ITT), mean difference -0.204, 95% CI -0.579 to 0.171 (PP)). These findings were wholly below the standardised non-inferiority limit of 0.4, suggesting that OST is non-inferior to CBT. No between-group differences were found on secondary outcomes. OST marginally decreased mean service use costs and maintained similar mean Quality Adjusted Life Years compared to CBT. CYP, their parents and the therapists found the intervention acceptable. Conclusions: OST has similar clinical effectiveness to CBT for specific phobias in CYP and maybe a cost-saving alternative.

***Keywords*—** one session therapy (OST), CBT, phobias, RCT.

Katie Biggs is with the University of Sheffield, United Kingdom (e-mail: c.e.biggs@sheffield.ac.uk).

Ceramic Surface Treatment through Laser Methods for Dental Application

Adil Othman Abdullah

***Abstract*—** Object: This study aimed to compare the impact of different laser scanning with that of conventional methods on zirconia surface treatment through evaluation of shear bond strength (SBS) values. Method: One hundred and thirty-two sintered zirconia cubic samples were prepared and randomly divided into six study groups: milling control (without surface treatment); grinding; sandblasting; and three-times, four-times, and five-times laser scanning groups. The treatment process for the first three groups was performed before the zirconia coating, while the last three groups were treated after zirconia coating with veneer slurry through a spraying technique. In the current study, the surface roughness Ra, contact angle measurement, phase transformation, topography and interfaces, SBS in unaged and aged conditions, and fracture mode patterns of zirconia cores were investigated. The results were analyzed using laser confocal scanning microscopy, drop analyzer, X-ray diffractometry (XRD), scanning electron microscope (SEM) equipped with energy dispersive spectroscopy (EDS), universal testing machine, and stereomicroscope. Results: The results indicated that three-times laser-scanned specimens presented higher Ra values than the other studied groups. The minimum contact angle value was detected in the mentioned group, while the control group presented the highest value. The XRD showed phase transformation from tetragonal to monoclinic t–m following grinding and sandblasting. However, the laser-scanned specimens and the control group preserved the structural integrity of the zirconia core, presenting the tetragonal phase only. The highest SBS values were recorded in specimens treated with three-times laser scanning in the unaged and aged conditions. A mixed fracture was a common fracture pattern among the studied groups. The results confirmed that SBS could be optimized through three-times laser scanning, and it provided better adhesion between zirconia and the veneer ceramic material. Conclusion: Multiple scanning processes of more than three times are not recommended for zirconia surface treatment.

***Keywords*—** ceramic, surface treatment, laser, contact angle.

Adil Othman Abdullah is with the Dental Department, Erbil Polytechnic University, Erbil, Kurdistan Region, Iraq (e-mail: dr.adil20@gmail.com).

Adil Othman Abdullah is with the Endodontics Department, Tishk International University, Erbil, Kurdistan Region, Iraq.

Adil Othman Abdullah is with the Conservative Department, Al-kitab University, Kirkuk, Kurdistan Region, Iraq.

Long-Term Opioid Therapy: Efficacy and Incidence of Substance Use Disorder

Julianne Macmullen

MCPHS University

Nur 820 Translational Research Practicum I

Dr. Adams

October 24, 2021

Author Note

This paper was reviewed by Rachel Schneider on 11/02/21 through MCPHS Writing Center online. Rachel Schneider reviewed this paper and discussed changes to sentence structure, APA format, and grammar.

Abstract

In the United States, many individuals suffer from substance use disorder (SUD) resulting from prescription therapy with opioids. This literature review examined the following question: For adults aged 25-60 with chronic non-cancer pain, does long-term use of opioids increases the risk of opioid abuse over one month? The methods used to examine this research question included an evaluation of recent, primary, peer-reviewed journal articles in the discipline of nursing and medicine found by utilizing MCPHS University's online library of databases. A critique and synthesis of the 10 articles chosen were then completed. The results of this research found two effective screening tools and prescription drug monitoring programs that providers can use to assess for SUD before prescribing opioids. This research also found that non-opioid pain medications may be as effective or more effective than long-term opioid therapy in treating chronic non-cancer pain. A final result found additional safety concerns, such as increased mortality rates, and the importance of information sharing and shared decision-making between the provider and patient. Gaps in the literature included limited populations and patient self-reported outcomes. Future research should be directed to examine the validity of patient self-reported data regarding prescription opioid drug use, motivations for opioid misuse, and pain-related outcomes over time. Implications for practice included low support for the initiation of long-term opioid prescription therapy for chronic non-cancer pain. This research also had implications for optimizing provider access to prescription drug monitoring programs.

Keywords: substance use disorder (SUD), opioids, long-term opioid therapy, chronic non-cancer pain

Long-Term Opioid Therapy: Efficacy and Incidence of Substance Use Disorder

The opioid epidemic affects over 10 million people a year in the United States alone and is now considered a national crisis (Substance Abuse, 2020). Over three-fourths of those who misuse opioids were initially prescribed opioids by their provider. Many individuals suffer from chronic pain and do not understand addiction risks when prescribed opioids (Klimas et al., 2019). Providers are given several screening tools and guidelines to follow when prescribing and monitoring the use of prescription opioids. The Centers for Disease Control and Prevention (CDC) provides specific guidelines for providers to use when prescribing opioids. Even with these guidelines, the number of opioid prescriptions continues to rise annually, with more than one-third of all United States adults prescribed opiates in 2015 (Gomes et al., 2017; Han et al., 2017; Liebschutz et al., 2017). These statistics make it necessary to research whether long-term opioid therapy, longer than one month, increases the risk of opioid misuse, even in individuals who were not at risk initially.

The purpose of this study is to educate providers on the risks of long-term opioid therapy and help providers identify individuals at increased risk of abusing opioids through clinical guidelines and evidence-based practice. This paper will also identify tools that providers can utilize to gather pertinent information, including risk factors, to determine the best treatment options for each patient. Finally, this paper will analyze the safety and efficacy of receiving long-term prescription opioid therapy in patients with low to no risk factors of opioid misuse to determine the risk of SUD in individuals prescribed opioids longer than one month. This research project also has implications for future practice to include new knowledge on the risk versus benefit of initiating long-term opioid therapy to patients with chronic non-cancer pain. As primary care providers, nurse practitioners (NPs) must have the tools and strategies necessary to identify patients at increased risk of drug misuse before prescribing opioids for pain. Much research has been conducted to examine the risk factors and provide prescription-specific guidelines and tools for the use of short-term (two to five days) opioid prescription therapies. However, the problem is that there is a gap in knowledge regarding whether long-term opioid therapy improves chronic non-cancer pain and what risk factors exist for SUD with opiate therapy prescribed longer than one month. This research project will examine the following question: For adults aged 25-60 with chronic non-cancer pain, does long-term use of opioids increase the risk of opioid abuse over one month?

Methodology

This literature review was achieved by an evaluation of recent, primary, peer-reviewed journal articles found by utilizing the MCPHS University's online library of databases. These databases included PubMed, EBSCO, eScholarship, Directory of Open Access Journals, and Gale Academic OneFile. The journal articles used in this study included articles published within the last five years from peer-reviewed, primary sources in the discipline of nursing and medicine. Keywords used to generate the journal articles included substance use disorder, long-term opioid therapy, chronic non-cancer pain, and opioids. The articles were evaluated through a review of each article's abstract to ensure the appropriateness of the article. The articles were then further evaluated by reviewing their population, methods, strengths, limitations, findings, and conclusions

Inclusion criteria included articles that had a study population between the ages of 25 and 60 who were prescribed long-term opioid therapy. Articles that were older than 5 years, did not include this adult population, or were not peer-reviewed journal articles were excluded from this study. In addition, only articles that involved long-term prescription opioid therapy and SUD due to long-term opioid therapy were included in this study. There were many articles discussing short-term opioid therapy and SUD and many other articles examining SUD due to mental health conditions. These articles were excluded from this study. Exclusion criteria also included any studies involving palliative care or cancer pain. From the articles included in this literature review, 10 studies were chosen to create a literature matrix for further evaluation. The articles included in the literature matrix are quantitative studies (Carey et al., 2018; Han et al., 2017; Häuser et al., 2020; Kang et al., 2019; Krebs et al., 2018; Larochelle et al., 2016; Morasco et al., 2020; Turner et al., 2019; Worley et al., 2017; Zgierska et al., 2016).

Theoretical Framework

The Theory of Bureaucratic Caring by Marilyn Anne Ray was used to guide this research project. This nursing theory gives a holistic view of caring, encompassing the ethical, spiritual, and cultural aspects of patient care (Smith & Parker, 2020). However, unlike other theories, this theory also emphasizes advocacy and the political, economic, legal, and technological factors of an organization. Ray's theory explains how all of these parts are interconnected and dynamic, forming a synchronized, holistic network of caring, communication, and relationships. Her theory also describes what happens when one part, such as cost and profit politics, becomes more valued than other parts, such as patient safety and ethics, causing a lack of confidence and trust in the health care system. Ray's theory explains that this lack of confidence is due to "health care systems falling victim to the corporatization of human enterprise" (Smith & Parker, 2020, p. 453). Finally, her theory describes bureaucracy as a living, ever-changing system of caring.

In this research project, Ray's theory was used to investigate and understand the opioid crisis and determine whether there is an increased incidence of substance use disorder (SUD) in individuals with no risk factors who are prescribed long-term opioid therapy for chronic non-cancer pain. This theory aided in examining and exploring caring as it pertains to the bureaucracy of opioid prescriptions and chronic non-cancer pain. In 2017, the Trump administration declared the opioid crisis a public health emergency, causing policymakers and stakeholders to create programs, laws, and policies to provide a solution to the opioid epidemic. The Controlled Substance Act is one form of politics set forth to regulate the distribution of opioids. This policy affected healthcare professionals more significantly than many other stakeholders, causing instability and unpredictability of the holistic health care system (Vranken et al., 2019). The Controlled Substance Act is one example of Ray's theory showing how the entire organization is affected when one part of the organization is affected.

Research studies found that a large number of patients treated with opioid prescriptions had a high morbidity and mortality rate due to SUD. In addition, marketing companies, pharmaceutical companies, and even some providers were found to be making a significant profit from opioid prescriptions, knowingly increasing the mortality and risk of SUD. Opioid market spending, pharmaceutical spending to physicians, and mortality rates of prescription opioid overdoses were examined from 2013 through 2015. The results directly correlated prescription opioid overdoses and pharmaceutical marketing to physicians (Hadland et al., 2019). This increase in the influence of pharmaceutical companies caused the health care system to become inequitable and misleading, producing a health care system that was no longer holistic or trusted by the public.

For these reasons, it is essential to research the efficacy of long-term opioid treatment for chronic non-cancer pain and examine whether long-term opioid prescriptions increase the risk of SUD even in the absence of other risk factors. Research must also be conducted to determine the most effective screening tools providers can use to rule out individuals at high risk of SUD for long-term opioid prescription therapy. Ray's theoretical framework, which blends the patient-centered, holistic caring for the patient with the organizational, economic, and political aspects of care, guided this research project in a balanced, systematic, and comprehensive manner. This framework allowed for the continued questioning and research to facilitate the healthcare system's transformation into a holistic environment of caring (Smith & Parker, 2020). Ray's framework also guided this research project by recognizing the interconnectedness of each aspect of the health care system to advocate for the best interests of both the patient and the system as a whole.

Critique of the Literature

This study utilized peer-reviewed journal articles published within the past five years. After evaluating roughly 60 articles, 10 articles were found to be sufficient to construct a literature matrix (Carey et al., 2018; Han et al., 2017; Häuser et al., 2020; Kang et al., 2019; Krebs et al., 2018; Larochelle et al., 2016; Morasco et al., 2020; Turner et al., 2019; Worley et al., 2017; Zgierska et al., 2016). Each article was inspected using Coughlan et al. (2007) and Ryan et al. (2007) to evaluate the article's strengths and weaknesses systematically. An article's usefulness is determined by ensuring that the elements influencing believability and the elements influencing the robustness of the research are adequate (Coughlan et al., 2007; Ryan et al., 2007). This critique evaluated each article's writing style, authors, report title, abstract, purpose, research problem, logical consistency, literature review, theoretical framework, aims, objectives, research question, hypotheses, sample, ethical considerations, operational definitions, methodology, data collection, instrument, design, validity, reliability, data analysis, results, discussion, conclusion, recommendations, and references. The purpose of this critique is to determine the quality of the articles used in this literature review and to ensure they are adequate to use in practice.

Critique of Quantitative Articles

Elements Influencing Believability of the Research

Writing Style. A research article's writing style should be well-written, concise, and grammatically correct (Coughlan et al., 2007). The researchers avoided the use of jargon and presented their research in an organized style that clearly labeled each section. As a result, there was an adequate flow to each article, which promotes unambiguousness in an understandable, easy-to-read manner (Carey et al., 2018; Han et al., 2017; Häuser et al., 2020; Kang et al., 2019; Krebs et al., 2018; Larochelle et al., 2016; Morasco et al., 2020; Turner et al., 2019; Worley et al., 2017; Zgierska et al., 2016). Keywords were clearly defined, providing transparency in four articles (Häuser et al., 2020; Kang et al., 2019; Morasco et al., 2020; Turner et al., 2019). The other six articles did not include keywords (Carey et al., 2018; Han et al., 2017; Krebs et al., 2018; Larochelle et al., 2016; Worley et al., 2017; Zgierska et al., 2016).

Authors. It is important that the researchers have adequate qualifications and knowledge of the research field (Coughlan et al., 2007). In the articles included in the literature matrix, the number of authors ranged from three to nine. All the article's researchers had sufficient knowledge in medicine, nursing, pharmacology, or psychiatry and earned a title of M.D., PhD, MA, MPH, or PharmD (Carey et al., 2018; Han et al., 2017; Häuser et al., 2020; Kang et al., 2019; Krebs et al., 2018; Larochelle et al., 2016; Morasco et al., 2020; Turner et al., 2019; Worley et al., 2017; Zgierska et al., 2016).

Report Title. The title of a research article should be clear and unambiguous to help direct practitioners and other researchers to the article (Coughlan et al., 2007). All of the articles adequately met these criteria. Six of the articles provided greater depth and transparency by indicating in the title the type of study to be completed (Han et al., 2017; Häuser et al., 2020; Kang et al., 2019; Morasco et al., 2020; Turner et al., 2019; Zgierska et al., 2016). An article's title should be between 10 and 15 words to be clear and concise (Coughlan et al., 2007). Five of the articles met this criterion (Carey et al., 2018; Kang et al., 2019; Larochelle et al., 2016; Morasco et al., 2020; Worley et al., 2017). Three of the articles longer than 15 words were also adequate and stated the aim, purpose, and type of study being done in the title (Han et al., 2017; Häuser et al., 2020; Kang et al., 2019). The other two article titles were excessively long and wordy (Krebs et al., 2018; Turner et al., 2019).

Abstract. An abstract offers the reader an overview of the purpose, aim, problem, methodology, results, and recommendations. The abstract also outlines the relevance to medical research and the clinical relevance of the study while also adequately defining the sample and the significant results of the study. An abstract that contains these elements provides a quick, comprehensive summary to inform the reader of the study being conducted without the need to read the entire article (Coughlan et al., 2007). Eight of the articles included these elements in their abstracts (Carey et al., 2018; Han et al., 2017; Kang et al., 2019; Krebs et al., 2018; Larochelle et al., 2016; Morasco et al., 2020; Worley et al., 2017; Zgierska et al., 2016). One article included each element except for recommendations in its abstract (Häuser et al., 2020). One article's abstract did not contain the findings, sample size, or recommendations (Turner et al., 2019).

Elements Influencing Robustness of the Research

Purpose/Research. Having a clearly stated, focused research problem is crucial to an adequate study (Coughlan et al., 2007). The research purpose is clearly stated in the abstract and background of each article (Carey et al., 2018; Han et al., 2017; Häuser et al., 2020; Kang et al., 2019; Krebs et al., 2018; Larochelle et al., 2016; Morasco et al., 2020; Turner et al., 2019; Worley et al., 2017; Zgierska et al., 2016). Two articles had a broad and overreaching purpose, leading to several limitations (Han et al., 2017; Worley et al., 2017). However, the researchers for all of the articles provided adequate support to demonstrate that their studies were significant to the medical, nursing, and pharmacology fields (Carey et al., 2018; Han et al., 2017; Häuser et al., 2020; Kang et al., 2019; Krebs et al., 2018; Larochelle et al., 2016; Morasco et al., 2020; Turner et al., 2019; Worley et al., 2017; Zgierska et al., 2016).

Problem. Identifying an unambiguous research problem is essential in guiding the research and filling a particular gap in knowledge (Coughlan et al., 2007). Each study clearly stated a research problem and adequately portrayed a gap in knowledge in a particular area. Each article also outlined how their research problem was crucial to their study and the medical field (Carey et al., 2018; Han et al., 2017; Häuser et al., 2020; Kang et al., 2019; Krebs et al., 2018; Larochelle et al., 2016; Morasco et al., 2020; Turner et al., 2019; Worley et al., 2017; Zgierska et al., 2016). Three of the studies' problems were overreaching, leading to limitations and a requirement for more research (Han et al., 2017; Turner et al., 2019; Worley et al., 2017).

Logical Consistency. A research study should have a logical consistency with links between each section in a sensible manner to provide a sound, coherent article (Coughlan et al., 2007). Each of the research articles adequately progressed between the steps of the research process, with clear links between each step. The articles began with the research purpose and significance for clinical practice and continued through the literature review, theoretical framework, methods, results, discussion, and recommendations (Carey et al., 2018; Han et al., 2017; Häuser et al., 2020; Kang et al., 2019; Krebs et al., 2018; Larochelle et al., 2016; Morasco et al., 2020; Turner et al., 2019; Worley et al., 2017; Zgierska et al., 2016).

Literature Review. A literature review is crucial to a research study to provide a depth of knowledge regarding the research problem and question as well as the gap in knowledge that is to be studied. The literature review provides a context to the area being studied and presents previous studies that correlate with the strengths and limitations of the study being conducted (Coughlan et al., 2007). The literature review of seven of the articles is logically organized and offers an unbiased, critical analysis of previous studies (Carey et al., 2018; Häuser et al., 2020; Kang et al., 2019; Morasco et al., 2020; Turner et al., 2019; Worley et al., 2017; Zgierska et al., 2016). Two articles are short and do not provide an in-depth review of the literature (Han et al., 2017; Krebs et al., 2018).

The majority of the literature reviews consisted of recent, relevant, primary empirical resources, which are peer-reviewed medical journal articles. Sources that are older than 10 years are included in the literature review because these sources are crucial due to the lack of research and the fact that they are still relevant to current practice (Carey et al., 2018; Han et al., 2017; Häuser et al., 2020; Kang et al., 2019; Krebs et al., 2018; Larochelle et al., 2016; Morasco et al., 2020; Turner et al., 2019; Worley et al., 2017; Zgierska et al., 2016). Eight of the literature reviews adequately provided the current knowledge as well as the gap in knowledge related to the purpose and problem of the study (Carey et al., 2018; Häuser et al., 2020; Kang et al., 2019; Larochelle et al., 2016; Morasco et al., 2020; Turner et al., 2019; Worley et al., 2017; Zgierska et al., 2016). Two of the articles did not provide an adequate background of knowledge related to their research problem and purpose (Han et al., 2017; Krebs et al., 2018).

Theoretical Framework. A theoretical framework should be identified and defined to give the research study clear boundaries and to guide the study (Coughlan et al., 2007). The theoretical framework or conceptual model was included in each article (Carey et al., 2018; Han et al., 2017; Häuser et al., 2020; Kang et al., 2019; Krebs et al., 2018; Larochelle et al., 2016; Morasco et al., 2020; Turner et al., 2019; Worley et al., 2017; Zgierska et al., 2016). Three of the articles used grounded theory to guide their research (Carey et al., 2018; Han et al., 2017; Morasco et al., 2020). Each of the other seven articles used different, clearly defined conceptual models to guide their research (Häuser et al., 2020; Kang et al., 2019; Krebs et al., 2018; Larochelle et al., 2016; Turner et al., 2019; Worley et al., 2017; Zgierska et al., 2016). The frameworks in each article were appropriate to their study (Carey et al., 2018; Han et al., 2017; Häuser et al., 2020; Kang et al., 2019; Krebs et al., 2018; Larochelle et al., 2016; Morasco et al., 2020; Turner et al., 2019; Worley et al., 2017; Zgierska et al., 2016).

Aims/Objectives/Research Question/Hypotheses. The aims, objectives, research question, and hypothesis should form a clear link and provide adequate direction to the study (Coughlan et al., 2007). The aims, objectives, and research question have been clearly identified in each article (Carey et al., 2018; Han et al., 2017; Häuser et al., 2020; Kang et al., 2019; Krebs et al., 2018; Larochelle et al., 2016; Morasco et al., 2020; Turner et al., 2019; Worley et al., 2017; Zgierska et al., 2016). Three of the articles have a clearly stated hypothesis (Krebs et al., 2018; Morasco et al., 2020; Worley et al., 2017). The other seven articles did not include a hypothesis (Carey et al., 2018; Han et al., 2017; Häuser et al., 2020; Kang et al., 2019; Larochelle et al., 2016; Turner et al., 2019; Zgierska et al., 2016). This is a predictable representation of a descriptive study, such as these studies, where the researcher refers to the purpose of the study or the research problem to guide their research (Coughlan et al., 2007).

Sample. The sample and sample size are crucial factors in a study's ability to represent the population being studied (Coughlan et al., 2007). Each of the articles clearly stated their sample, target population, and sample size (Carey et al., 2018; Han et al., 2017; Häuser et al., 2020; Kang et al., 2019; Krebs et al., 2018; Larochelle et al., 2016; Morasco et al., 2020; Turner et al., 2019; Worley et al., 2017; Zgierska et al., 2016). Three articles included inclusion and exclusion criteria for their sample population (Häuser et al., 2020; Larochelle et al., 2016; Turner et al., 2019). All the articles included a target population of adults older than 18 years (Carey et al., 2018; Han et al., 2017; Häuser et al., 2020; Kang et al., 2019; Krebs et al., 2018; Larochelle et al., 2016; Morasco et al., 2020; Turner et al., 2019; Worley et al., 2017; Zgierska et al., 2016). The sample size was adequate in each study, and the sample selection was clearly defined (Carey et al., 2018; Han et al., 2017; Häuser et al., 2020; Kang et al., 2019; Krebs et al., 2018; Larochelle et al., 2016; Morasco et al., 2020; Turner et al., 2019; Worley et al., 2017; Zgierska et al., 2016). The sample population was narrow in three articles, causing limitations in either geographic location or sex distribution (Kang et al., 2019; Krebs et al., 2018; Worley et al., 2017).

Ethical Considerations. Ethical considerations that must be considered in a research study include autonomy, non-maleficence, beneficence, and justice to ensure a study is conducted in an ethical manner. Each study should also include an ethical committee or institutional review board to determine whether the study is ethical and will give approval for their study (Coughlan et al., 2007). The participants were fully informed about the nature of the research, and autonomy and confidentiality were ensured in each study (Carey et al., 2018; Han et al., 2017; Häuser et al., 2020; Kang et al., 2019; Krebs et al., 2018; Larochelle et al., 2016; Morasco et al., 2020; Turner et al., 2019; Worley et al., 2017; Zgierska et al., 2016). Ethical permission by either an ethical committee or institutional review board was approved in each study. The participants were protected from harm in nine of the studies (Carey et al., 2018; Han et al., 2017; Häuser et al., 2020; Kang et al., 2019; Larochelle et al., 2016; Morasco et al., 2020; Turner et al., 2019; Worley et al., 2017; Zgierska et al., 2016). One of the studies provided a potential for harm as the participants in the sample were prescribed opioid and non-opioid medications (Krebs et al., 2018).

Operational Definitions. Any terms, theories, or concepts must be clearly defined to ensure the reader has a clear understanding of the study (Coughlan et al., 2007). All terms, theories, and concepts are clearly defined throughout each article (Carey et al., 2018; Han et al., 2017; Häuser et al., 2020; Kang et al., 2019; Krebs et al., 2018; Larochelle et al., 2016; Morasco et al., 2020;

Turner et al., 2019; Worley et al., 2017; Zgierska et al., 2016).

Methodology. The methodology of the research design and the data gathering instrument should be clearly labeled and defined in the methods section of an article (Coughlan et al., 2007). The research design is clearly defined in each article (Carey et al., 2018; Han et al., 2017; Häuser et al., 2020; Kang et al., 2019; Krebs et al., 2018; Larochelle et al., 2016; Morasco et al., 2020; Turner et al., 2019; Worley et al., 2017; Zgierska et al., 2016). Three articles are quasi-experimental studies (Krebs et al., 2018; Morasco et al., 2020; Worley et al., 2017). Seven of the articles are non-experimental (Carey et al., 2018; Han et al., 2017; Häuser et al., 2020; Kang et al., 2019; Larochelle et al., 2016; Turner et al., 2019; Zgierska et al., 2016).

Data Collection. The data collection method should be clearly defined and adequate for the research study (Coughlan et al., 2007). Five of the studies' research designs were conducted through a review of electronic health records (Carey et al., 2018; Han et al., 2017; Häuser et al., 2020; Larochelle et al., 2016; Morasco et al., 2020;). Five studies used an itemized survey questionnaire (Kang et al., 2019; Krebs et al., 2018; Turner et al., 2019; Worley et al., 2017; Zgierska et al., 2016). Each article clearly stated and described the data gathering instrument in the methods section (Carey et al., 2018; Han et al., 2017; Häuser et al., 2020; Kang et al., 2019; Krebs et al., 2018; Larochelle et al., 2016; Morasco et al., 2020; Turner et al., 2019; Worley et al., 2017; Zgierska et al., 2016).

Instrument Design. The design of a study's instrument should be clearly defined and elicit accurate information to adequately achieve the goals of the study (Coughlan et al., 2007). The instrument design is clearly depicted for the reader, and the researchers established that the instrument was appropriate to support each studies' aims and objectives by adequately measuring the outcomes (Carey et al., 2018; Han et al., 2017; Häuser et al., 2020; Kang et al., 2019; Krebs et al., 2018; Larochelle et al., 2016; Morasco et al., 2020; Turner et al., 2019; Worley et al., 2017; Zgierska et al., 2016). Each study utilized a previously designed instrument and described why that instrument was the most appropriate for their study.

Validity and Reliability. The validity and reliability of an instrument are crucial to the study's believability (Coughlan et al., 2007). For an instrument to be valid, it must measure what it is meant to measure, and for an instrument to be reliable, it must consistently and accurately measure what it is supposed to measure (Coughlan et al., 2007). Each article utilized a well-established instrument and explained how that instrument met the validity and reliability requirements (Carey et al., 2018; Han et al., 2017; Häuser et al., 2020; Kang et al., 2019; Krebs et al., 2018; Larochelle et al., 2016; Morasco et al., 2020; Turner et al., 2019; Worley et al., 2017; Zgierska et al., 2016). Seven of the articles linked their instruments to other studies conducted with similar results or with similar areas of interest (Han et al., 2017; Häuser et al., 2020; Kang et al., 2019; Krebs et al., 2018; Turner et al., 2019; Worley et al., 2017; Zgierska et al., 2016). Three articles directly described and explained their instrument using relevant research and statistics to show why it was appropriate for the study (Carey et al., 2018; Larochelle et al., 2016; Morasco et al., 2020).

Data Analysis and Results. A research article should clearly include the statistical analysis and why the specific tests were chosen and used. An article should also reveal the results of the tests and statistics clearly and concisely (Coughlan et al., 2007). Each article clearly stated their data analysis and results as well as described why each test was chosen and how it correlated with their results and research problem (Carey et al., 2018; Han et al., 2017; Häuser et al., 2020; Kang et al., 2019; Krebs et al., 2018; Larochelle et al., 2016; Morasco et al., 2020; Turner et al., 2019; Worley et al., 2017; Zgierska et al., 2016). Greater than 50% of each article's sample participated in the studies. Each article clearly defined the significance of the findings.

Discussion/Conclusion/Recommendations. The discussion, conclusion, and recommendations should flow logically and relate to the literature review and purpose of the study (Coughlan et al., 2007). Each article's findings were linked back to the literature review, and each of the article's strengths and weaknesses, including generalizability, were clearly discussed (Carey et al., 2018; Han et al., 2017; Häuser et al., 2020; Kang et al., 2019; Krebs et al., 2018; Larochelle et al., 2016; Morasco et al., 2020; Turner et al., 2019; Worley et al., 2017; Zgierska et al., 2016). A recommendation for future research was clearly stated in each article. Three articles have a clearly stated hypothesis identified in the discussion section of the articles (Krebs et al., 2018; Morasco et al., 2020; Worley et al., 2017). Of these three articles, two identified their hypothesis as supported in their discussion (Morasco et al., 2020; Worley et al., 2017). One article's hypothesis was not supported, which was clearly explained in the discussion section (Krebs et al., 2018)

References. An accurate list of any books, journal articles, reports, or other media should be included in the reference list. The reference list should include mostly primary peer-reviewed articles published within the past five years of the study's publication date (Coughlan et al., 2007). Each article accurately referenced the journals, reports, books, or other media (Carey et al., 2018; Han et al., 2017; Häuser et al., 2020; Kang et al., 2019; Krebs et al., 2018; Larochelle et al., 2016; Morasco et al., 2020; Turner et al., 2019; Worley et al., 2017; Zgierska et al., 2016). Six of the articles included 50% or greater of their references published within the past five years (Carey et al., 2018; Han et al., 2017; Häuser et al., 2020; Larochelle et al., 2016; Turner et al., 2019; Zgierska et al., 2016 et al., 2016). Three articles did not contain any references published longer than 10 years ago (Carey et al., 2018; Häuser et al., 2020; Larochelle et al., 2016). One article contains 10 references published longer than 10 years ago, which is three times the number of other articles (Morasco et al., 2020).

Synthesis of the Literature

A synthesis of the literature was completed to compare the themes in each study. Common themes arising from the literature review included screening tools clinicians can use to assess for SUD, prescription drug monitoring programs, information sharing, Nonopioid pain medication options, and safety risks for long-term opioid therapy. Much research has shown the benefits of short-term opioid prescription therapy for acute pain events. However, there is no efficient research to conclude that long-term opioid therapy effectively reduces chronic pain. There is also a lack of research regarding the risks of long-term opioid therapy for chronic non-cancer pain and its association with SUD. Research has found that the majority of overdoses occurred through prescription opioid medications obtained by a provider (Carey et al., 2018 et al., 2018; Han et al., 2017; Häuser et al., 2020; Larochelle et al., 2016; Turner et al., 2019; Zgierska et al., 2016). A significant number of providers continued to prescribe long-term opioid therapy even after an overdose (Carey et al., 2018; Larochelle et al., 2016). This synthesis of the literature demonstrated the lack of efficacy of long-term opioid prescription therapy and its potential to cause SUD.

Screening Tools

Each patient should be assessed before the initiation of long-term opioid therapy to ensure the risk of SUD is low. Several studies mentioned misuse patterns clinicians can observe and assess before beginning opioid therapy (Carey et al., 2018; Larochelle et al., 2016; Morasco et al., 2020). These misuse patterns included the utilization of more than one pharmacy or more than one provider to obtain opioid prescriptions (Carey et al., 2018; Larochelle et al., 2016). Risky behaviors were another misuse pattern defined in the literature as a potential pattern of misuse (Carey et al., 2018; Morasco et al., 2020). In addition, two studies offered specific screening tools: addiction behavior checklists and the National Institute of Drug Abuse drug screening tool to assess for the potential of SUD Before prescribing opioid prescription medications (Krebs et al., 2018; Larochelle et al., 2016).

Prescription Drug Monitoring

Another important aspect of long-term opioid treatment for chronic non-cancer pain was prescription drug monitoring (Carey et al., 2018; Kang et al., 2019; Larochelle et al., 2016). Monitoring is often completed by state-run databases known as prescription drug monitoring programs. These programs collect data on all opioids prescribed from pharmacies in their state and allow providers to view this information for each patient (Carey et al., 2018; Larochelle et al., 2016). A few states even mandate that providers review the data in each patient's database before prescribing opioids to the patient (Carey et al., 2018; Larochelle et al., 2016). These measures help to reduce the number of prescriptions a single patient can obtain at any given time. In addition, these programs also allow providers to review the patient's risk for opioid misuse by allowing examination of the patient's opioid prescription history (Carey et al., 2018; Larochelle et al., 2016). Prescription drug monitoring can also be accomplished through pharmacy-physician collaboration (Carey et al., 2018; Kang et al., 2019).

Information Sharing and Shared Decision Making

Information sharing leads to improved care and outcomes (Kang et al., 2019; Larochelle et al., 2016; Turner et al., 2019). Public awareness of the opioid crisis has led to many policies and procedures for prescribing opioid medications for chronic non-cancer pain. Studies show that even with heightened public awareness of the opioid epidemic, many individuals did not understand the risks of long-term prescription opioid use (Kang et al., 2019; Turner et al., 2019). Five of the studies discussed the need for the provider to inform the patient of the risks of prescription opioid therapy and its potential for SUD before prescribing them (Carey et al., 2018; Häuser et al., 2020; Kang et al., 2019; Larochelle et al., 2016; Turner et al., 2019).

Shared decision-making is also discussed in four studies (Häuser et al., 2020; Kang et al., 2019; Larochelle et al., 2016; Turner et al., 2019). Keeping open communication and a good rapport with the patient was crucial in patient-centered care (Kang et al., 2019; Turner et al., 2019). The research found that 63% of adults in the U.S. who have abused prescription opioids received the prescription for pain relief from a provider (Han et al., 2017; Larochelle et al., 2016; Turner et al., 2019). Of those with SUD in the U.S., 40% of opioids were acquired through a friend or family member with a prescription from a provider (Han et al., 2017; Turner et al., 2019). These statistics showed that physical pain is the most common reason for prescription opioid use and that many opioid prescriptions were prescribed in excess of what the patient consumed, creating the potential for unused opioids to be abused (Han et al., 2017; Larochelle et al., 2016; Turner et al., 2019).

Non-opioid Pain Medications

NSAIDs and other non-opioid pain medications and therapies were options that were also effective at managing chronic non-cancer pain in many individuals (Han et al., 2017; Häuser et al., 2020; Krebs et al., 2018; Morasco et al., 2020; Worley et al., 2017; Zgierska et al., 2016). Six studies showed that opioid medications did not demonstrate any advantage in pain management over non-opioid treatments (Han et al., 2017; Häuser et al., 2020; Krebs et al., 2018; Morasco et al., 2020; Worley et al., 2017; Zgierska et al., 2016). Three of these studies also suggested that non-opioid treatments provided better management of pain severity than opioid medications (Häuser et al., 2020; Krebs et al., 2018; Worley et al., 2017). In fact, two of the studies found that long-term opioid therapy had little benefit of pain relief when compared with placebo (Krebs et al., 2018; Worley et al., 2017).

The results of these studies suggested low support for the initiation of long-term opioid prescription therapy for chronic non-cancer pain (Häuser et al., 2020; Krebs et al., 2018; Worley et al., 2017). This information is crucial for providers to be aware that non-opioid medications and alternative therapies can be more beneficial to patients who suffer from chronic non-cancer pain. The results of these studies showed that higher doses of opioids did not improve pain management (Han et al., 2017; Häuser et al., 2020; Krebs et al., 2018; Morasco et al., 2020; Worley et al., 2017). These studies also showed that higher doses of opioids resulted in poorer outcomes and higher mortality rates (Han et al., 2017; Häuser et al., 2020; Krebs et al., 2018; Morasco et al., 2020; Worley et al., 2017). The same studies showed more medication-related adverse symptoms with opioid medications than with non-opioid medications (Han et al., 2017; Häuser et al., 2020; Krebs et al., 2018; Morasco et al., 2020; Worley et al., 2017). Overall, six of these studies suggested that long-term opioid therapy for chronic non-cancer pain provided poor pain management and higher functional impairment (Han et al., 2017; Häuser et al., 2020; Krebs et al., 2018; Morasco et al., 2020; Worley et al., 2017; Zgierska et al., 2016).

Safety

Prescription Dose Escalation

The safety of long-term opioid therapy for chronic non-cancer pain is another theme in this synthesis of the literature. Four studies indicated that a higher dose of opioid prescription therapy over a long period might increase pain and pain sensitivity while also increasing the patient's tolerance to opioids (Han et al., 2017; Häuser et al., 2020; Krebs et al., 2018; Morasco et al., 2020). This information is highly concerning, considering that to achieve long-term pain management, the patient would continue to require a higher dose of opioid medications, forming a vicious cycle of pain relief, opioid dependence, increased tolerance, and then heightened pain once again. There is a lack of evidence to suggest that a prescription dose increase will provide a beneficial effect to the patient; however, providers continue to prescribe higher doses to patients with chronic non-cancer pain (Han et al., 2017; Häuser et al., 2020; Krebs et al., 2018; Morasco et al., 2020). In addition, three studies examined the relationship between opioid dose escalations and changes in pain severity (Han et al., 2017; Häuser et al., 2020; Morasco et al., 2020). Each study's results were similar, showing little to no association with an increased dose of opioid prescription therapy and decreased pain severity (Han et al., 2017; Häuser et al., 2020; Morasco et al., 2020).

Mortality Rate

Another safety issue with long-term opioid therapy is that it causes a higher mortality rate than other individuals of the same age and disease process (Han et al., 2017; Häuser et al., 2020; Krebs et al., 2018; Morasco et al., 2020). Mortality continues to rise among individuals prescribed long-term opioid therapy as the prescription doses increase (Han et al., 2017; Häuser et al., 2020; Krebs et al., 2018; Morasco et al., 2020). The percent of individuals with mental health conditions also increases with long-term opioid prescription therapy (Han et al., 2017; Häuser et al., 2020; Morasco et al., 2020). The most common mental health disorders these patients suffered from included depression, anxiety, insomnia, and self-harm behaviors (Morasco et al., 2020; Quinn et al., 2017). It was found that the longer an individual was prescribed long-term opioid therapy, the more psychological conditions an individual suffered (Morasco et al., 2020; Quinn et al., 2017). Research also found that the risk of SUD and overdose affected those with mental health disorders greater than other individuals who did not have a history of mental health disorders (Krebs et al., 2018; Morasco et al., 2020; Quinn et al., 2017). This finding supports the need for continued mental health evaluation during long-term opioid prescription therapy.

Quality of life

Finally, there was an evident decline in quality of life for those who were prescribed long-term opioids for chronic non-cancer pain. This decline in quality of life is due to many factors. One factor includes the adverse effects of opioid pain medications such as sleep disturbance, hypothalamic-pituitary dysregulation, physical disability, constipation, depression, sedation, and addiction (Baldini et al., 2018; Morasco et al., 2020). These patients also experienced a decline in their quality of life due to their chronic pain. In three of the research studies in this literature review, evidence showed that there was little or no correlation with opioid prescription dose increases and better pain management and show a lower quality of life for those prescribed higher doses for more extended periods of time (Han et al., 2017; Häuser et al., 2020; Morasco et al., 2020). These studies also showed that long-term opioid prescriptions provided more risks than benefits to individuals with chronic non-cancer pain (Han et al., 2017; Häuser et al., 2020; Morasco et al., 2020).

Conclusion

This synthesis of the literature recommended that providers utilize screening tools and prescription drug monitoring programs when prescribing opioid medications to patients (Carey et al., 2018; Kang et al., 2019; Larochelle et al., 2016). If providers are prescribing long-term opioid prescriptions to patients with chronic non-cancer pain, it is suggested that providers evaluate patients for potential misuse patterns, discuss the risks of long-term opioid prescription therapy, counsel patients on the dangers of sharing or selling any unused opioids, and reevaluate each patient before refilling an opioid prescription for chronic non-cancer pain (Carey et al., 2018; Häuser et al., 2020; Kang et al., 2019; Larochelle et al., 2016; Turner et al., 2019). Patients should also be informed of long-term opioid prescription therapy risks, including increased mortality, decreased functional ability, and physiologic and psychologic dependence (Krebs et al., 2018; Han et al., 2017; Häuser et al., 2020; Morasco et al., 2020).

This synthesis of the literature also found that non-opioid prescriptions are as effective or more effective than long-term prescription opioid therapy in managing chronic non-cancer pain (Han et al., 2017; Häuser et al., 2020; Krebs et al., 2018; Morasco et al., 2020; Worley et al., 2017; Zgierska et al., 2016). Overall, long-term opioid therapy for chronic non-cancer pain provided poor pain management and higher functional impairment in this literature review (Han et al., 2017; Häuser et al., 2020; Krebs et al., 2018; Morasco et al., 2020; Worley et al., 2017; Zgierska et al., 2016). This literature review also exposed many safety concerns transpiring with long-term prescription opioid therapy. These safety concerns included an increased likelihood of mental health disorders with long-term prescription opioid use, higher mortality rates, lower quality of life, increased pain sensitivity, and increased tolerance to opioid medications (Han et al., 2017; Häuser et al., 2020; Krebs et al., 2018; Morasco et al., 2020).

Gaps in the Literature

A gap in the literature addresses an area of the research topic that has been understudied or unexplored. These gaps in the literature present areas for future research. Throughout the literature review, several gaps present potential limitations to this research topic. These gaps included data gathering techniques, population, sample size, research methods, and research variables.

Data Gathering

Three studies utilized information from insurance claims to gather their data (Carey et al., 2018; Han et al., 2017; Larochelle et al., 2016). However, data gathered using insurance claims is a potential limitation because it eliminates the individuals who do not use insurance but instead use cash for prescription medications (Carey et al., 2018). The utilization of insurance claims also excludes the most vulnerable populations who are at higher risk of a substance use disorder, such as the homeless, who may not have medical insurance (Carey et al., 2018; Han et al., 2017; Larochelle et al., 2016). Finally, insurance claims do not account for individuals who buy prescriptions drugs off the street, are given prescription drugs from family members or friends, or obtain prescription drugs by other means.

Population

Another gap in the literature included the population chosen for the studies. Seven of the studies used limited populations, which may not generalize to the entire population (Carey et al., 2018; Kang et al., 2019; Krebs et al., 2018; Larochelle et al., 2016; Morasco et al., 2020; Worley et al., 2017; Zgierska et al., 2016). Three of the studies included a population of Veterans Affairs patients and commercially insured individuals (Carey et al., 2018; Krebs et al., 2018; Larochelle et al., 2016). These results exclude any individual on Medicare or uninsured, which may underestimate the number of individuals with substance misuse disorder and subsequent overdose. Two studies were limited in geographic location (Kang et al., 2019; Zgierska et al., 2016). A study with a limited geographic location may not show results that are generalized to the overall population. Two studies were limited to patients in the emergency department or inpatient facilities only (Larochelle et al., 2016; Worley et al., 2017). The results of these studies exclude deaths before a medical encounter and individuals who do not receive medical care at an emergency department or inpatient setting but instead may be treated at an outpatient care facility and those who do not seek care. Five of the studies included a population with a majority of individuals being over the age of 50 (Carey et al., 2018; Häuser et al., 2020; Krebs et al., 2018; Morasco et al., 2020; Zgierska et al., 2016). These studies' results may not be accurate for younger individuals. Five of the studies included a White majority population in their sample (Carey et al., 2018; Han et al., 2017; Krebs et al., 2018; Morasco et al., 2020; Worley et al., 2017; Zgierska et al., 2016). These studies may not be generalized to minority populations. Four of the studies did not investigate demographic information in their sample population (Häuser et al., 2020; Kang et al., 2019; Larochelle et al., 2016; Turner et al., 2019). Demographic data is valuable information to indicate which population is represented in the study.

Sample Size

A third gap in the literature included a small sample size. Three of the studies included small sample sizes as a limitation to their study (Kang et al., 2019; Worley et al., 2017; Zgierska et al., 2016). A sample size that is too small may increase the margin of error for the study. A small sample may also not represent the general population adequately. When a sample is too small, randomization for replication of the results may be altered, causing nonreproducible results (Yan et al., 2017).

Research Methods

A fourth gap in the literature involved the research methods used in the studies. All ten of the research studies were quantitative studies (Carey et al., 2018; Han et al., 2017; Häuser et al., 2020; Kang et al., 2019; Krebs et al., 2018; Larochelle et al., 2016; Morasco et al., 2020; Turner et al., 2019; Worley et al., 2017; Zgierska et al., 2016). Quantitative studies are clinically useful in providing noncausal relationships between variables. However, quantitative studies do not establish casual relationships between variables. Qualitative studies provide information regarding opinions, feelings, and individual experiences. This type of information is valuable in assessing substance use disorder, overdose, chronic non-cancer pain, and the side effects and benefits of long-term opioid use to provide better insight into an individual's motivations for misuse.

Research Variables

Finally, the different research variables presented gaps in the literature. Four of the studies involved patient self-reported outcomes (Han et al., 2017; Krebs et al., 2018; Morasco et al., 2020; Zgierska et al., 2016). Self-reported outcomes have not been validated for individuals with opioid misuse (Han et al., 2017; Krebs et al., 2018; Morasco et al., 2020; Zgierska et al., 2016). Results formed from self-reported data form a potential reporting bias favoring opioids. Another research variable included the possibility of the ICD-9-CM codes to misclassify overdose (Häuser et al., 2020; Larochelle et al., 2016). ICD-9-CM codes are generated administratively into a database where data was obtained for two studies (Häuser et al., 2020; Larochelle et al., 2016). Misclassifications of overdose may have resulted in an underestimation of overdose events and all-cause mortality. A final gap in the literature included studies that were unable to identify whether individuals used their medications as prescribed, saved medications for later use, gave medications away, or never used them at all (Häuser et al., 2020; Krebs et al., 2018). This lack of information underestimated the number of prescription opioids left unused after a prescription was filled.

Implications for Future Research

Future research should be directed to examine the validity of patient self-reported data regarding prescription opioid drug use (Han et al., 2017; Krebs et al., 2018; Morasco et al., 2020; Zgierska et al., 2016). Patient self-reported data can easily be manipulated or include a patient's bias toward the information being ascertained. Several studies utilized patient self-reported data, making it necessary to research its validity to support these studies. It would also be beneficial for future studies to evaluate the motivations for opioid misuse (Carey et al., 2018; Han et al., 2017; Larochelle et al., 2016; Morasco et al., 2020). As providers begin to understand the motivations to misuse opioids, providers can evaluate patients for specific misuse patterns before prescribing opioid medications. Research on motivations for misuse would also provide valuable data to form patient questionnaires and other screening tools providers can use to evaluate patients prescribed long-term opioid treatment.

Another area future studies should research is the physiologic opioid dependence in patients with long-term use of prescription opioids (Häuser et al., 2020; Krebs et al., 2018). Physiologic dependence is an essential aspect of opioid dose escalations for pain control. This type of dependence makes it difficult for providers to discontinue the opioid medication due to patient withdrawal symptoms and rebound pain events. Insight into the physiologic dependence of long-term opioid prescriptions would provide valuable data into methods providers can utilize to reduce withdrawal symptoms and rebound pain events. Future research should also examine pain-related outcomes over time (Häuser et al., 2020; Morasco et al., 2020; Worley et al., 2017). These outcomes would help practitioners to form guidelines for best practices in prescribing long-term opioid therapy.

Future research should also examine ways to inform communities of the risks of long-term opioid treatment (Turner et al., 2019; Worley et al., 2017). The short-term risks of opioids have been effectively communicated to the community through television commercials, providers, social media, and many other forms of communication delivery. Unfortunately, many individuals have not been advised of the risks associated with long-term opioid therapy, such as increased mortality rates, lower functional ability, and increased risk of substance misuse, and physiologic and psychologic dependence. Informing the public of these increased risks may help to decrease the number of individuals prescribed long-term opioid therapy and allow individuals to make informed decisions on their health.

It would also be beneficial for future studies to examine clinical and collaborative interventions that improve the safety of individuals at high risk of substance use disorder (Carey et al., 2018; Kang et al., 2019; Krebs et al., 2018; Larochelle et al., 2016; Zgierska et al., 2016). Additional research on prescription drug monitoring programs may be beneficial in improving patient safety and provider awareness of patient misuse patterns. Further research into other forms of pain control, such as alternative therapies, non-opioid pain medications, and development into new effective therapies, may enhance patient safety and lower mortality rates for chronic non-cancer pain management (Zgierska et al., 2016). Another area that would improve patient safety that researchers should examine is how providers counsel their patients after an overdose (Larochelle et al., 2016; Zgierska et al., 2016). It is unknown whether providers who prescribe opioids after an overdose know of the overdose event and whether the provider conducts additional counseling after these events occur. It would also be beneficial to form practice guidelines to know what types of counseling are the most effective in reducing opioid overdose to enhance patient safety.

Implications for Practice

Multidisciplinary Treatment

This synthesis of the literature provided several implications for practice. One implication for practice involved a combination of treatment modalities that would reduce pain severity (Han et al., 2017; Kang et al., 2019; Zgierska et al., 2016). These modalities included a combination of multidisciplinary treatment options for patients with chronic non-cancer pain who require long term therapy and included alternative therapies, behavioral therapies, non-opioid pain medications, opioid pain medications, psychotherapy, pain management clinics, and physical therapy (Han et al., 2017; Häuser et al., 2020; Kang et al., 2019; Larochelle et al., 2016; Turner et al., 2019; Worley et al., 2017; Zgierska et al., 2016).

Information Sharing

Another implication for practice involved reducing opioid misuse in the community through providers informing the public of the most recent data regarding the risks of long-term prescription opioid therapy for chronic non-cancer pain (Han et al., 2017; Häuser et al., 2020; Larochelle et al., 2016; Turner et al., 2019; Worley et al., 2017). Providers should be aware of opioid prescription selling, sharing, and diversions and mediate these risks with patient education and counseling by informing patients of the risks and keeping open communication to reduce the risk of substance use disorder and all-cause mortality (Han et al., 2017; Häuser et al., 2020; Larochelle et al., 2016; Morasco et al., 2020; Worley et al., 2017). Much research suggested widespread availability of prescription opioids indicative of prescription under-usage (Han et al., 2017; Häuser et al., 2020; Larochelle et al., 2016; Morasco et al., 2020; Worley et al., 2017). With knowledge of this information, providers can better assess the needs of their patients by inspecting and recording how many opioids are left in each prescription before refilling the prescription (Han et al., 2017). It may also be beneficial to discontinue automatic refills and instead refill opioid prescriptions on an as-needed basis (Han et al., 2017).

Prescription Dose Escalation

A third implication for practice is the low support for the initiation of long-term opioid prescription therapy for chronic non-cancer pain (Häuser et al., 2020; Krebs et al., 2018; Worley et al., 2017). Increasing the dose of a prescription opioid for chronic non-cancer pain did not improve pain sensitivity or functioning (Han et al., 2017; Häuser et al., 2020; Morasco et al., 2020;

Worley et al., 2017). However, higher doses of opioids resulted in poorer outcomes and higher mortality rates (Han et al., 2017; Häuser et al., 2020; Morasco et al., 2020; Worley et al., 2017). Providers should attempt to decrease opioid prescribing to reduce the number of unused opioid medications available for misuse (Han et al., 2017).

Misuse Patterns

A final implication for future practice was for providers to monitor for misuse patterns on an ongoing basis throughout opioid therapy (Carey et al., 2018; Han et al., 2017; Larochelle et al., 2016; Morasco et al., 2020). Two of the studies had implications for practice for optimizing prescription drug monitoring programs for more straightforward interpretation and availability to providers (Carey et al., 2018; Han et al., 2017). Prescription drug monitoring programs may have the ability to form an interface that operates in conjunction with electronic health records to monitor for multiple prescriptions and multiple prescribers and alert providers of potential misuse patterns (Carey et al., 2018).

Conclusion

With many individuals prescribed opioids to relieve chronic pain, it is essential to research the efficacy and risks of long-term opioid treatment. It is also crucial to understand an individual's risk factors and which tools are available to providers to screen for opioid abuse and decrease the risk of SUD. This research project conducted a literature review to examine whether long-term opioid therapy increases the risk of SUD in individuals with chronic non-cancer pain. This review of literature provided essential, up-to-date knowledge on the risks and benefits of long-term opioid prescription therapy for chronic non-cancer pain. It has also provided valuable screening tools, information, and recommendations on prescription drug monitoring programs, and patterns for misuse that providers can evaluate before prescribing opioid therapy. It has been well known that opioid pain medications are effective when treating acute pain events. However, one essential result found from this research included non-opioid medications as effective or more effective than long-term opioid therapy at treating chronic non-cancer pain. Another result of this research found additional safety concerns for long-term prescription opioids, such as increased mortality rates, lower functional ability, decreased quality of life, and increased pain sensitivity and tolerance to opioid medications. A final result of this research included the importance of information sharing and shared decision-making between the provider and the patient.

Throughout the literature review, several gaps present potential limitations to this research topic. These gaps included data gathering techniques, population, sample size, research methods, and research variables. Future research should be directed to examine the validity of patient self-reported data regarding prescription opioid drug use, motivations for opioid misuse, and pain-related outcomes over time. Future studies would also benefit from researching physiologic opioid dependence in patients with long-term use of prescription opioids. Future research should also examine ways to inform communities of the risks of long-term opioid treatment along with the clinical and collaborative interventions that improve the safety of individuals at high risk of substance use disorder and how and when providers counsel patients after an overdose. Finally, future research should include other forms of pain control, such as alternative therapies, non-opioid pain medications, and the development of new effective therapies that may enhance patient safety and lower mortality rates for chronic non-cancer pain management.

This research provided several implications for practice. Implications for practice included low support for the initiation of long-term opioid prescription therapy for chronic non-cancer pain. Another implication for practice involved a combination of treatment modalities which be more beneficial in reducing pain severity. This research also had implications for optimizing prescription drug monitoring programs for easier interpretation and availability to providers. Another implication for practice involved reducing opioid misuse in the community through providers informing the public of the most up-to-date data regarding long-term prescription opioid therapy risks for chronic non-cancer pain. A final implication for future practice was for providers to monitor for misuse patterns on an ongoing basis throughout opioid therapy.

References

Baldini, A. G., Von Korff, M., & Lin, E. H. (2018). A review of potential adverse effects of long-term opioid therapy. *The Primary Care Companion For CNS Disorders*. https://doi.org/10.4088/pcc.11m01326

Carey, C. M., Jena, A. B., & Barnett, M. L. (2018). Patterns of Potential Opioid Misuse and Subsequent Adverse Outcomes in Medicare, 2008 to 2012. *Annals of Internal Medicine*, *168*(12), 837. https://doi.org/10.7326/m17-3065

Coughlan, M., Cronin, P., & Ryan, F. (2007). Step-by-step guide to critiquing research. Part I: Quantitative research. British Journal of Nursing, 16(2), 658-663.

Gomes, T., Pasricha, S., Martins, D., Greaves, S., Tadrous, M., Bandola, D., Singh, S., Paterson, M., Mamdani, M., & Juurlink, D. (2017). Behind the prescriptions: a snapshot of opioid use across all Ontarians. https://doi.org/10.31027/odprn.2017.04

Hadland, S. E., Rivera-Aguirre, A., Marshall, B. D., & Cerdá, M. (2019). Association of Pharmaceutical Industry Marketing of Opioid Products With Mortality From Opioid-Related Overdoses. *JAMA Network Open*, *2*(1). https://doi.org/10.1001/jamanetworkopen.2018.6007

Han, B., Compton, W. M., Blanco, C., Crane, E., Lee, J., & Jones, C. M. (2017). Prescription Opioid Use, Misuse, and Use Disorders in U.S. Adults: 2015 National Survey on Drug Use and Health. *Annals of Internal Medicine*, *167*(5), 293. https://doi.org/10.7326/m17-0865

Häuser, W., Schubert, T., Vogelmann, T., Maier, C., Fitzcharles, M.-A., & Tölle, T. (2020). All-cause mortality in patients with long-term opioid therapy compared with non-opioid analgesics for chronic non-cancer pain: a database study. *BMC Medicine*, *18*(1). https://doi.org/10.1186/s12916-020-01644-4

Kang, I., Urick, B., Vohra, R., & Ives, T. J. (2019). Physician-pharmacist collaboration on chronic non-cancer pain management during the opioid crisis: A qualitative interview study. *Research in Social and Administrative Pharmacy*, *15*(8), 1027–1031. https://doi.org/10.1016/j.sapharm.2019.04.052

Klimas, J., Gorfinkel, L., Fairbairn, N., Amato, L., Ahamad, K., Nolan, S., Simel, D. L., & Wood, E. (2019). Strategies to Identify Patient Risks of Prescription Opioid Addiction When Initiating Opioids for Pain. *JAMA Network Open*, *2*(5). https://doi.org/10.1001/jamanetworkopen.2019.3365

Krebs, E. E., Gravely, A., Nugent, S., Jensen, A. C., DeRonne, B., Goldsmith, E. S., Kroenke, K., Bair, M. J., & Noorbaloochi, S. (2018). Effect of Opioid vs Nonopioid Medications on Pain-Related Function in Patients With Chronic Back Pain or Hip or Knee Osteoarthritis Pain. *JAMA*, *319*(9), 872. https://doi.org/10.1001/jama.2018.0899

Larochelle, M. R., Liebschutz, J. M., Zhang, F., Ross-Degnan, D., & Wharam, J. F. (2016). Opioid Prescribing After Nonfatal Overdose and Association With Repeated Overdose. *Annals of Internal Medicine*, *165*(5), 376. https://doi.org/10.7326/l16-0168

Liebschutz, J. M., Xuan, Z., Shanahan, C. W., LaRochelle, M., Keosaian, J., Beers, D., Guara, G., O'Connor, K., Alford, D. P., Parker, V., Weiss, R. D., Samet, J. H., Crosson, J., Cushman, P. A., &amp; Lasser, K. E. (2017). Improving Adherence to Long-term Opioid Therapy Guidelines to Reduce Opioid Misuse in Primary Care. JAMA Internal Medicine, 177(9), 1265. https://doi.org/10.1001/jamainternmed.2017.2468

Morasco, B. J., Smith, N., Dobscha, S. K., Deyo, R. A., Hyde, S., & Yarborough, B. J. (2020). Outcomes of prescription opioid dose escalation for chronic pain: results from a prospective cohort study. *Pain*, *161*(6), 1332–1340. https://doi.org/10.1097/j.pain.0000000000001817

Quinn, P. D., Hur, K., Chang, Z., Krebs, E. E., Bair, M. J., Scott, E. L., Rickert, M. E., Gibbons, R. D., Kroenke, K., & D'Onofrio, B. M. (2017). Incident and long-term opioid therapy among patients with psychiatric conditions and medications: A national study of commercial health care claims. *Pain*, *158*(1), 140–148. https://doi.org/10.1097/j.pain.0000000000000730

Ryan, F., Coughlan, M., & Cronin, P. (2007). Step-by-step guide to critiquing research. Part 2: Qualitative research. British Journal of Nursing, 16(22), 738-744

Seth, P., Rudd, R. A., Noonan, R. K., & Haegerich, T. M. (2018). Quantifying the Epidemic of Prescription Opioid Overdose Deaths. *American Journal of Public Health*, *108*(4), 500–502. https://doi.org/10.2105/ajph.2017.304265

Shah, A., Hayes, C. J., &amp; Martin, B. C. (2017). Characteristics of Initial Prescription Episodes and Likelihood of Long-Term Opioid Use — United States, 2006–2015. MMWR. Morbidity and Mortality Weekly Report, 66(10), 265–269. https://doi.org/10.15585/mmwr.mm6610a1

Smith, M. C., & Parker, M. E. (2020). *Nursing theories and nursing practice* (5th ed). F.A. Davis. pp. 449-468.

Substance Abuse and Mental Health Services Administration. (2020). *Key substance use and mental health indicators in the United States: Results from the 2019 National Survey on Drug Use and Health* (HHS Publication No. PEP20-07-01-001, NSDUH Series H-55). Rockville, MD: Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration. Retrieved from https://www.samhsa.gov/data/

Turner, J. P., Caetano, P., & Tannenbaum, C. (2019). Leveraging policy to reduce chronic opioid use by educating and empowering community dwelling adults: a study protocol for the TAPERING randomized controlled trial. *Trials*, *20*(1). https://doi.org/10.1186/s13063-019-3508-z

Vranken, M. J. M., Linge-Dahl, L., Mantel-Teeuwisse, A. K., Radbruch, L., Schutjens, M.-H. D. B., Scholten, W., Payne, S., & Jünger, S. (2019). The perception of barriers concerning opioid medicines: A survey examining differences between policy makers, healthcare professionals and other stakeholders. *Palliative Medicine*, *34*(4), 493–503. https://doi.org/10.1177/0269216319894190

Worley, M. J., Heinzerling, K. G., Shoptaw, S., & Ling, W. (2017). Volatility and change in chronic pain severity predict outcomes of treatment for prescription opioid addiction. *Addiction*, *112*(7), 1202–1209. https://doi.org/10.1111/add.13782

Yan, F., Robert, M., & Li, Y. (2017). Statistical methods and common problems in medical or biomedical science research. International journal of physiology, pathophysiology and pharmacology, 9(5), 157–163.

Zgierska, A. E., Burzinski, C. A., Cox, J., Kloke, J., Stegner, A., Cook, D. B., Singles, J., Mirgain, S., Coe, C. L., & Bačkonja, M. (2016). Mindfulness Meditation and Cognitive Behavioral Therapy Intervention Reduces Pain Severity and Sensitivity in Opioid-Treated Chronic Low Back Pain: Pilot Findings from a Randomized Controlled Trial. *Pain Medicine*, *17*(10), 1865–1881. https://doi.org/10.1093/pm/pnw006

Community Re-Integrated Soldiers' Perceptions of Barriers and Facilitators to A Home-Based Physical Rehabilitation Programme Following Lower-Limb Amputation

Ashan Wijekoon, Abi Beane, Subashini Jayawardana

***Abstract*—** Background: Soldiers' physical rehabilitation and long term health status has been hindered due to limited investment in and access to rehabilitation services. Home-based rehabilitation programmes could offer a potentially feasible alternative to facilitate long-term recovery. Objectives: To explore Sri Lankan soldiers' perceptions of barriers and facilitators to a home-based physical rehabilitation programme.Methods and Materials: We conducted qualitative semi-structured interviews with community re-integrated army veterans who had undergone unilateral lower limb amputation following war related trauma. Veterans were identified from five districts of Sri Lanka, based on a priori knowledge of veteran community settlements (Disabled Category Registry) obtained from Directorate of Rehabilitation, MoD, Sri Lanka. Individuals were stratified for purposive selection. The interview guide was developed from existing methods and adapted for context. Verbatim transcripts of interviews were analyzed for emerging themes using an inductive approach. Following consent, participants met the researcher (AW- a trained physiotherapist fluent in Sinhalese). Results: Twenty-five Interviews were conducted, totaling 7.2 hours of new data (Mean±SD: 0.28±0.11). All participants were male, aged 30-55 years (Mean±SD: 46.1±7.4), and had experienced traumatic amputation as a result of conflict. Twenty-four sub themes were identified. Inadequate space for exercises, absence of equipment and assistance to conduct the exercises at home, alongside absence of community healthcare services were all barriers. Burden of comorbidities, including chronic pain and disability level, were also barriers. Social support systems, including soldier societies, family, and kinship with other amputees, were seen as facilitators to an at-home programme. Motivation for independence was a strong indicator of engagement. Conclusion: Environment, chronic pain, and absence of well-established community health services were key barriers. Family and soldier support was a facilitator. Engagement with community healthcare providers (physiotherapist and primary care physicians) will be essential to the success of an at-home rehabilitation program.

***Keywords*—** physical rehabilitation, home-based, soldiers, disability, lower-limb amputation, qualitative.

Ashan Wijekoon is with Department of Allied Health Sciences, Faculty of Medicine, University of Colombo, Sri Lanka, (e-mail: ashan@med.cmb.ac.lk).

Abi Beane is with Nuffield Department of Clinical Medicine, University of Oxford, United Kingdom.

Subashini Jayawardana is with Department of Allied Health Sciences, Faculty of Medicine, University of Colombo, Sri Lanka.