National Conference on

RECENT TRENDS IN HUMANITARIAN TECHNOLOGY

NCRTHT 2025

Oct 30-31, 2025

Marian Engineering College
Kazhakuttam-Thiruvananthapuram

In collaboration with

NISH

National Institute of Speech & Hearing

THIRUVANANTHAPURAM

NATIONAL CONFERENCE ON
RECENT TRENDS IN HUMANITARIAN TECHNOLOGY

IN COLLABORATION WITH

NISH

NATIONAL INSTITUTE OF SPEECH AND HEARING

THIRUVANANTHAPURAM

OCTOBER (30 – 31) 2025

MARIAN ENGINEERING COLLEGE

KAZHAKUTTAM, THIRUVANANTHAPURAM, KERALA

NATIONAL CONFERENCE ON
RECENT TRENDS IN HUMANITARIAN TECHNOLOGY

COMMITTEES:

General Chair:

1. Dr. Abdul Nizar M , Principal MEC
2. Dr. Manoj M, Prof. MEC

TECHNICAL PROGRAM COMMITTEE

TPC Chair

1. Dr. Shanthi K J, Prof. MEC
2. Ms. Vinitha B Elza, HOD Dept. ECE MEC

Organizing Secretary:

1. Subha P S, Assoc. Prof. MEC
2. Ramola P Joy, Assoc. Prof. MEC

Members:

1. Arya Manoharan, NISH, Thiruvananthapuram
2. Amith G Nair, NISH, Thiruvananthapuram
3. Nitturi Naresh Kumar, IBSC AMTZ Visakapatanam
4. Dr. Ajish K Abraham, Professor AIISH, Mysore
5. Dr. Arun Sasidharan, Scientist, NIMHANS,Bangalore
6. Dr. D. Vaithiyanathan, NIT Delhi
7. Dr. Ravish, Professor, Dr. AIT Bangalore
8. Dr. Leela Rengaraj, Professor, NIT Trichi

Patrons:

1. Rev. Dr. A. R. John, Manager – MEC
2. Dr. A Samson, Dean - Academics, MEC
3. Rev. Fr. Jim Roch, Bursar - MEC

ISBN NO. 978-93-92321-65-8

Published by
Owl Books,
Thiruvananthapuram

Marian Engineering College

The Marian Engineering College, with its campus at Trivandrum, Kerala, India, was established in the year 2001 and is affiliated to the APJ Abdul Kalam Technological University. The college is approved by All India Council for Technical Education, New Delhi and offers Undergraduate and post graduate degree courses in Engineering, which are accredited by National Board of Accreditation (NBA). The College is being managed by the Trivandrum Social Service Society (TSSS) which is under the control of Trivandrum Latin Catholic Archdiocese. His grace Most Rev. Dr. M.Soosa Pakiam, Arch Bishop of Trivandrum is the chief patron of this college. The college is well equipped and the students are assisted by a team of highly qualified, experienced and talented teachers who constantly strive for academic excellence.

Department of Electronics & Communication

The Department of Electronics & Communication of Marian Engineering College was started in the year2001. The department has been successful in training more than 1000 students to become Electronics Engineers of high capability and credibility, in every stream around the globe. We are proud of all our students, alumni and teachers for working hand-in-hand to make the Electronics venture a success with constant excellence in academics and effective Industry- Institute Interaction.

Department Vision

The Department of Electronics & Communication Engineering to be recognized at national & international level by creating engineers capable of accepting the challenges of ever-changing technologies for the betterment of the mankind.

Department Mission

To promote a rigorous learning environment for both theory and practice, and help the students develop professional and communication skills To foster research, innovation & entrepreneur skills in faculty and students for the benefit of society

CONTENT

Addressing Internal Team Challenges in Engineering Management Organizations Through System Dynamics Modeling

1st Dr. Khumbelo Difference Muthavhine
Department of Electrical Engineering
South Africa Government
Tshwane, South Africa
kdmuthavhine@gmail.com

***Abstract*—Internal team challenges are the concerns of most engineering project managers who struggle during the project's progress. Most managers agreed that System Dynamics (SD) modeling could help with internal team problems; however, most project managers detest SD modeling because of its complexity, particularly when they don't have the necessary computation, software development, IT, or engineering skills. The study analyzed internal team concerns using experimental and literature review approaches. Literature review results from studies show that 78.3333% to 95% of project managers find it difficult to deal with internal team issues. Managers may utilize SD to apply mathematical skills to internal team difficulties, as evidenced by a range of 78.3333% to 95.3%; however, developing an SD model lacking programming and mathematical analytics knowledge is challenging. In response to these difficulties, the authors developed a model SD model that may assist managers in resolving issues on internal team obstacles.**

***Index Terms*—Internal team challenges; SD model, Project, Management, Engineering**

I. Introduction

Internal team challenges are the typical problems that engineering management organizations encounter with no practical solutions [1]– [7]. These issues must be effectively managed, fixed, and resolved. Problems that emerge within a project team are referred to as internal team challenges in engineering management. These issues arise from interpersonal conflicts amongst team members [6]- [7]. Among these issues include a lack of enthusiasm, a lack of accountability, a lack of expertise, a lack of organizational direction, oversight that throws the team out of cooperation, and inadequate or poor communication [1]– [7].

Internal team challenges as a tendency problem for management are thought to be fundamental to the human condition, they are bound to occur, especially in the dynamic workplace with hierarchical structure and complicated care issues and problems. This article's goal is to illustrate the possible advantages of managing internal team difficulties under constructive team leadership. Resolving internal team issues has several advantages, such as strengthening relationships inside the team, improving personal relationships, creating multiple solutions for each issue, and removing unfavorable effects. Refer to Figure 1.

Many managers and literature reviews employ the conventional (traditional) methods shown in Figure 2 to solve the internal team challenges. Additionally, the purpose of this article is to improve the traditional approaches by introducing a novel approach to system dynamics modeling, which has been disregarded by several managers and literature reviews for a long time. This article gives the reasons why several managers and literature reviews ignore system dynamics.

A. Internal Team Challenges

Internal team challenges in engineering management refer to concerns that develop within a project team [1]- [7]. These issues arise as a result of interpersonal conflicts among team members. These issues include poor (or insufficient) communication, a lack of strategic direction, supervision that leads to team misalignment, a lack of enthusiasm, a lack of accountability, and a lack of expertise [1]- [7]. The dynamic team-building interventions are crucial for increasing internal communication in engineering management [24]. Eliminating communication barriers and cultivating a healthy communication atmosphere can increase cooperation, collaboration, and results [24]. Engineering management should incorporate dynamic team-building interventions into their overall organizational development plan [24]. engineering managers must strike a balance between technical expertise and interpersonal abilities to lead different teams and guarantee that all participants work together towards agreed project goals [25].

Poor leadership and team dynamics are susceptible to communication breakdowns, disputes, and decreased team performance, negatively impacting project results. To solve these problems, engineering managers should build strong

Identify applicable funding agency here. If none, delete this.

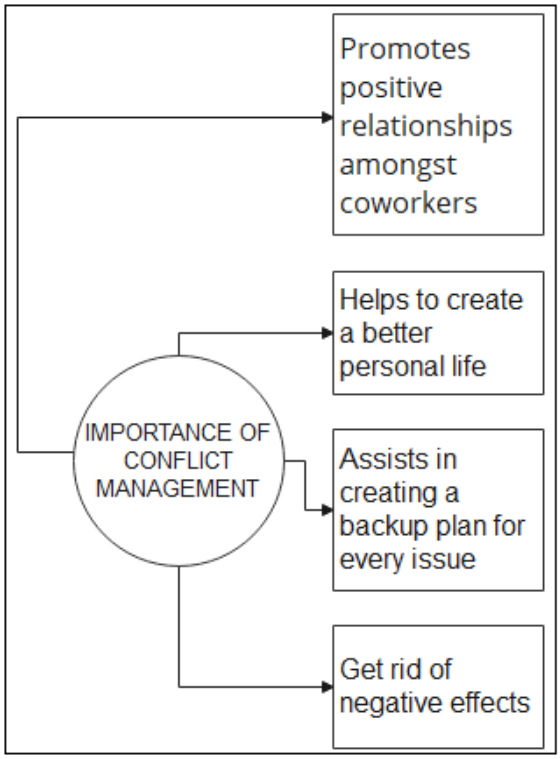


Fig. 1. Benefits of Resolving Conflicts in Internal Team [9]

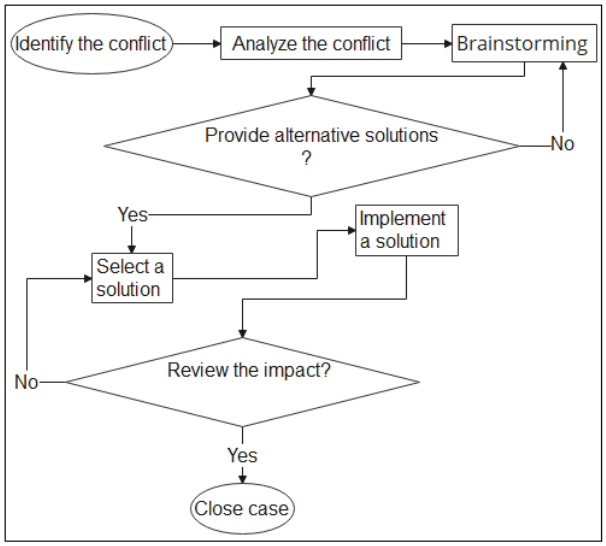


Fig. 2. Standard and Common Model of Solving Internal Team Conflict Management [83]

leadership abilities, establish effective communication channels, and foster a healthy team culture [25]. Engineering managers should foster an open and trusting culture, encouraging team members to share ideas, concerns, and feedback [25]. According to Htet *et al.* [25], engineering managers must have conflict resolution skills because disputes can develop from different sources, such as differing ideas, conflicting objectives, and cultural differences. Managing teamwork challenges can boost productivity and morale [26]. Understanding common workplace teaming challenges and successful solutions can help resolve arguments and problems [26].

Teamwork challenges vary by industry or area, but some frequent ones include lack of clarity, trust concerns, personality conflicts, withholding information, lack of communication, decreased involvement, excessive staff numbers, internal competitiveness, philosophical differences, habitual confrontations, opposing aims, working alone, lack of self-awareness, and skill overlap [26].

Internal teams can become socio-emotionally stressful, posing obstacles that can harm mutual trust and shared mental models [27].

To address the issues with internal teams, many managers and literature reviews use the typical (traditional) approaches depicted in Figure 2. This article aims to enhance conventional methods by proposing a new technique for modeling system dynamics, which has been neglected for a while by several managers and literature assessments. This article explains why system dynamics are ignored by several managers and literature studies.

The authors of this paper recommend adopting System Dynamics (SD) as an alternative to conventional methods for addressing internal team challenges due to the paucity of research on the subject. In addition, SD has several advantages over traditional techniques such as the Critical Path method, PERT, Kanban Board, Gantt Chart, and Timeline [16]– [19]. The writers selected system dynamics above other traditional instruments because of the following advantages:

i. SD was first intended to resemble industrial dynamics. Ever since, there has been noticeable progress in the understanding, evaluation, administration, and financial uses of SD. According to [82].
ii. Project managers ought to be able to conquer these obstacles and make sense of complicated phenomena by applying an SD methodology along with computerized system models. As stated in [80].
iii. SD can be used to learn more about the elements, behaviors, and possible impacts of different circumstances and policies on intricate systems. As stated in [81].
iv. The scientific education sectors should be exposed to SD as a cross-cutting concept that has promise for solving a wide range of societal issues. As stated in [80].
v. By breaking a system down into interconnected chains of stocks and flows that interact with one another through both positive and negative feedback loops, SD may be used to examine and evaluate a system's behavior over time. As stated in [82].
vi. When evaluating theories regarding the origins and consequences of opportunities or systemic issues, SD may be a helpful technique. As stated in [81].
vii. Large, intricate, interconnected structures can be handled more easily because of SD features, which also lessen time and effort requirements, much like project management challenges in general. As stated in [82].

viii. SD is a method that may be applied to a variety of disciplines and fields to aid in the understanding of complicated phenomena and the resolution of difficult problems. As stated in [80].
ix. SD can be utilized for evaluating various scenarios and methods for innovation in systems or development, and it can then provide stakeholders and decision-makers with data and recommendations. As stated in [81].
x. With the SD technique, the dynamic interactions found inside entire systems may be looked at and studied. As stated by [82].
xi. The power of the SD technique lies in its capacity to replicate the actions and output of a system without requiring the system to undergo pre-analysis and evaluation. As stated in [82].
xii. SD may be a beneficial teaching and learning tool for cognitive functions. as recommended by [81].
xiii. Establishing an SD offers several benefits. Initially, by revealing the essential framework of the system, SD has the potential to streamline and enhance the clarity of intricate circumstances. Second, by advancing our understanding of the dynamics and architecture of complex systems, SD may make it possible to create extremely successful long-term transformation strategies. Furthermore, through the development of planned models and decision simulators, SD could support policy-making and education. As stated in [82]. Despite all of SD's advantages, little has been attempted to use SD to resolve problems inside teams.

The causes of why many project managers are hesitant to use SD solutions to solve management problems are discussed in the section that follows.

B. Why Don't Most Project Managers Use SD to Address Internal Team Issues?

Due to SD complexity, most project managers despise SD modeling, particularly if they lack experience in programming mathematical analysis, engineering, or IT [28], [30], [34]. SD modeling requires computer science, causal loop diagrams, models, looping feedback, delays, programming mathematically, analytical decision-making, stock comprehension, handling flaws, elements of control, compute models, feedback from information theory [84], [29], [31], [35]. Please refer to Figure 2 for a summary explanation. The authors of this article address the internal team issues associated with SD. Thus, the description of the problem. The problem statement for the study is provided in the subsection that follows.

II. Problem Statement

Internal team issues plague engineering management companies, making it challenging to find workable solutions [1]–[7]. These problems need to be successfully handled, corrected and remedied. In engineering management, internal team difficulties are issues that arise within a project team. Interpersonal disputes amongst team members cause these problems [6]-[7]. These problems include a lack of passion, a lack of responsibilities, a lack of knowledge, a lack of organizational direction, oversight that causes misalignment in the team, and poor (or insufficient) communication [1]- [7].

III. Importance of the Study

After the study, the manager and community will benefit from the following underpinned results of this research and work: To find out if

IV. Research Objectives

The contribution of the study is as follows:

i. Understanding issues that project managers have with work assigned to internal teams.
ii. Understanding if project managers are knowledgeable about SD models, which can help with internal team problems during project management.
iii. Understanding if project managers know about the issues that develop inside their teams.
iv. Understanding if managers can create and build SD models using mathematical formulas and programming skills to address issues with their teams.
v. Understanding if there is an SD model that can address issues inside a team readily available for free download from the public domain (such as periodicals, little CDs, or Google).

V. Research Questions

This study's crux primary research question is: Do managers use IoT devices for management functions? The following are the secondary research questions:

i. Do project managers have issues with work assigned to internal teams?
ii. Are project managers knowledgeable about SD models, which can help with internal team problems during project management?
iii. Do project managers know about the issues that develop inside their teams?
iv. Can managers create and build SD models using mathematical formulas and programming skills to address issues with their teams?
v. Is an SD model that can address issues inside a team readily available for free download from the public domain (such as periodicals, little CDs, or Google)?

VI. Literature Review

There are three subsections in this section. The literature review on internal team challenges is the first subsection. The reason why the majority of project managers do not use SD to address internal team difficulties is covered in the second subsection. The benefits of employing SD are covered in the third subsection.

A. Internal Team Problems Literature Review

Internal team challenges in engineering management refer to concerns that develop within a project team [1]- [7]. These issues arise as a result of interpersonal conflicts among team members. These issues include poor (or insufficient) communication, a lack of strategic direction, supervision that leads to team misalignment, a lack of enthusiasm, a lack of accountability, and a lack of expertise [1]- [7]. Benarkuu and Katere [24] found that dynamic team-building interventions are crucial for increasing internal communication in engineering management. According to the findings of Benarkuu and Katere [24], dynamic team-building interventions conducted in the study organizations significantly improved internal communication. Benarkuu and Katere [24] suggest that dynamic team-building interventions can improve internal communication in engineering management. According to Benarkuu and Katere [24], eliminating communication barriers and cultivating a healthy communication atmosphere can increase cooperation, collaboration, and results. Engineering management should incorporate dynamic team-building interventions into their overall organizational development plan [24].

According to Htet $et\ al.$ [25], engineering management relies heavily on excellent communication and collaboration across team members to ensure project success. According to Htet $et\ al.$ [25], engineering managers must strike a balance between technical expertise and interpersonal abilities to lead different teams and guarantee that all participants work together towards agreed project goals. According to Htet $et\ al.$ [25], poor leadership and team dynamics are susceptible to communication breakdowns, disputes, and decreased team performance, negatively impacting project results. To solve these problems, engineering managers should build strong leadership abilities, establish effective communication channels, and foster a healthy team culture [25]. Engineering managers should foster an open and trusting culture, encouraging team members to share ideas, concerns, and feedback [25]. According to Htet $et\ al.$ [25], engineering managers must have conflict resolution skills because disputes can develop from different sources, such as differing ideas, conflicting objectives, and cultural differences.

According to Birt [26], teamwork issues are a natural aspect of managing staff, and competent managers understand how to recognize and solve them. Managing teamwork challenges can boost productivity and morale [26]. According to Birt [26], effective teamwork can boost departmental or corporate productivity. Understanding common workplace teaming challenges and successful solutions can help resolve arguments and problems [26]. According to Birt [26], teamwork challenges vary by industry or area, but some frequent ones include lack of clarity, trust concerns, personality conflicts, withholding information, lack of communication, decreased involvement, excessive staff numbers, internal competitiveness, philosophical differences, habitual confrontations, opposing aims, working alone, lack of self-awareness, and skill overlap.

According to Kazemitabar $et\ al.$ [27], internal teams can become socio-emotionally stressful, posing obstacles that can harm mutual trust and shared mental models. The study by Kazemitabar $et\ al.$ [27] aims to explore and classify general teamwork issues in a setting to identify challenges that inhibit the development of important teamwork mechanisms (i.e., trust among teammates and shared conceptual models). Kazemitabar $et\ al.$ [27] identified 16 general problems that hinder teamwork in an engineering environment. A model of team difficulties was created to divide obstacles into macro-level topics such as motivational, cognitive, emotional, and behavioral challenges.

Nasution and Bazin [34] believe that a project management plan assists decision-makers in managing project progress. It was outstanding in achieving results for its clients. Many governmental endeavors fail to achieve their goals [34]. The problem originates from insufficient problem mapping, political considerations, misunderstanding of regulations and rules by authorized entities, and insufficient monitoring to detect deviations [34].

Kazemitabar $et\ al.$ [27] identified the problems that prevented the establishment of trust between individuals and common mental models. Kazemitabar $et\ al.$ [27] findings provide vital insights for instructors and coaches in recognizing the types of teamwork issues that may develop in project management. The results also enlighten educators on which problems likely contribute to trust between individuals breakdown and weak shared psychological bonds [27].

Nasution and Bazin [34] propose leveraging SD, project management, and decision-making to successfully solve these difficulties. Nasution and Bazin [34] suggest that decision-making and project management are used to develop and specify SD equations and models. Project managers use simulation findings from SD models to handle and monitor their projects [34].

B. Why Most Project Managers are not Using SD for Internal Team Problems?

Many project managers were hesitant to utilize SD modeling on internal team problems due to its limitations [28]. According to Manenzhe, Telukdarie, and Munsamy [29], there were still unsolved concerns with the use of expert and programming mathematically for SD in project management on internal team problems. Rumeser and Emsley [28] identified three significant challenges: altering mental models, involving stakeholders, promoting change, and effectively explaining and applying the notion. According to Amin $et\ al.$ [30], there was a lack of literature review approaches and knowledge of dynamic capabilities, which rendered it impossible to investigate how SD talents might be exploited in management decision-making on internal team problems.

According to Eidin $et\ al.$ [80], project managers encountered difficulties with SD modeling due to eleven root causes that were found by using Ishikawa's fishbone approach. These difficulties can be divided into three primary groups: mental model changing, involving stakeholders and spearheading

modifications, and convincingly articulating and putting the model into practice. They were all associated with personnel management [80].

According to Ghaffarzadegan, Lyneis, and Richardson [35], modest SD models might be helpful for planning and internal team problems. Two models were studied and used as examples by Ghaffarzadegan, Lyoneis, and Richardson [35] to demonstrate how small SD models might be able to address the most pressing issues facing policymakers.

Ghaffarzadegan, Lyoneis, and Richardson [35] observed that small SD models had drawbacks. First, whether they were real or imagined, decision-makers and buyers alike often looked for a particular model that considered every conceivable chain of events. Decision-makers ran the risk of losing faith if they discovered that their anticipated variable or link was missing from the model in such scenarios. Stakeholders usually desire to have individual representations of their departments, businesses, or communities, which lead to more division, according to Ghaffarzadegan, Lyoneis, and Richardson [35]. It could be helpful in that situation to have a specific version of the model and demonstrate that the final behavior was still mostly dependent on the important connections or factors that the policymakers had predicted for internal team problems.

According to David and Margaret [81], there may be certain issues to be mindful of while applying system dynamics. It may take a lot of time and resources to develop and validate system dynamics models, and the models might not fully account for all the pertinent information and uncertainties in the system. Furthermore, the models may contain the modeler's prejudices and presumptions, which makes it challenging to understand and communicate with non-experts. Moreover, the approximations and simplifications of reality may cause resistance or distrust from particular audiences [81].

According to Rumeser and Emsley [28], many project managers were found unfamiliar with the importance of SD and lacked optimism regarding the model and internal team problems. Manenzhe, Telukdarie, and Munsamy [29] discovered that support operations' predicted value-adds to company profitability were ineffective and inefficient. Manenzhe, Telukdarie, and Munsamy [29] found that obstacles to SD competency were subjective and lacked rationality and decision-making, especially in high-production settings like internal team problems.

Shahfarzadegan, Lynneis, and Richardson [35] claimed that if the outcomes of a simpler model were well known, then a more complex model might be built to assess smooth decision outcomes for internal team problems. The results of Ghaffarzadegan, Lyneis, and Richardson [35] indicated that because of the limitations of tiny SD models, modelers might underestimate the importance of significant feedback loops in reality. Good micro SD models ought to be simple to understand and comprise all of the most fundamental loops for internal team problems.

According to Mashaly and Fernald [82], project managers' ignorance was the root of their troubles when it came to SD modeling. For example, applying the system dynamics technique could be a convenient and straightforward way to describe causation links among numerous variables and factors. However, because of basic models that could be fixated on pointless and irrelevant minutiae, such ease and convenience could lead to inaccurate outcomes. According to Mashaly and Fernald [82], to avert these cases through evaluations, causal link assessments, and causality checks, knowledge and experience are needed. The existence of causal linkages for validation and calibration procedures is determined with the aid of these tests and checks, which significantly raises the integrated modeling process's reliability.

According to Rumeser and Emsley [28], managers disliked using SD for a single project. According to Amin $et\ al.$ [30], project managers might enhance the utilization of computers by simulating the system under evaluation through computerized experiments. Amin $et\ al.$ [30] used computer models to simulate complicated occurrences. Models could not substitute real-world systems or internal team problems. Rumeser and Emsley [28] found that managers avoid using SD due to perceived impracticality.

Manenzhe, Telukdarie, and Munsamy [29] proposed SD as a potential solution for opinionated project management, control issues, and internal team problems. Manenzhe, Telukdarie, and Munsamy [29] found that efficient SD modeling improved comprehensive business performance.

It was suggested by Ghaffarzadegan, Lyoneis, and Richardson [35] that smaller models with fewer feedback loops would be simpler to build than bigger ones. Ghaffarzadegan, Lyoneis, and Richardson [35] had shown that compact SD models could only arise following a detailed examination of a larger model, allowing for the discovery and isolation of only the most prevalent feedback loops. Ghaffarzadegan, Lyoneis, and Richardson [35] demonstrated that after a larger model was constructed and the modeler identified the common loops, the project manager could create a smaller version to submit to policymakers of the internal team challenges. Furthermore, it had been shown by Ghaffarzadegan, Lyoneis, and Richardson [35] that the development of small SD models shouldn't obstruct modeling, or "operational reasoning".

Project managers had difficulties with SD modeling, according to Mashaly and Fernald [82], because validation could be crucial in solving problems related to internal team obstacles that could be solved by mathematical modeling in general and SD modeling in particular. Unfortunately, validation of SD models may be far more challenging than validation of a black box model (Mashaly and Fernald [82]). The internal model frame verification process may be extremely conceptually and technically complex, which was the rationale behind it.

Rumeser and Emsley [28] found that project managers often lacked trust in the SD model, accountability regarding past assumptions, and internal team problems. Models aimed to simplify and abstract issue situations to a specific level of the internal team problems.

According to Amin $et\ al.$ [30], a model's running step should resemble reality. If that was not the case, the problem should be identified and resolved. Models should be examined

based on their usability and appropriateness for the project challenge, rather than their ability to accurately represent reality. Manenzhe, Telukdarie, and Munsamy [29] discovered that optimizing maintenance plans to internal team problems efforts was a typical difficulty in project management.

Mashaly and Fernald [82] came to the conclusion that the internal model frame might present philosophical challenges due to its possible close relationship and direct connection to the central philosophical issue of determining the validity of scientific arguments. In addition, the challenge may be technically complex if established validations—like tests of a statistical hypothesis—do not exist to ascertain whether the model frame was, in fact, near the problem frame [82].

According to Ghaffarzadegan, Lyoneis, and Richardson [35], SD could promote thoughtful conversations on causality and how variables impact behavior. Modelers should clearly express the causal link between the variables [35]. According to Ghaffarzadegan, Lyneis, and Richardson [35], accurate definitions of causal links and clarification of significant capacity limits are essential components of a sound modeling technique. Even though all causal interactions might not be included in tiny models, variables should nonetheless be operational at a high level [35]. GHaffarzadegan, Lyneis, and Richardson [35] found that small SD models could nevertheless greatly enhance planning despite these drawbacks of internal team problems.

Manenzhe, Telukdarie, and Munsamy [29] indicated that managers might not embrace SD models depending on the engineering equipment used. According to Crookes $et\ al.$ [32], project management disliked SD due to the impracticality of examining all possible aspects and their relationships in internal team problems. According to Amin $et\ al.$ [30], project managers had misgivings about SD. To capture the foundations of a problem situation, researchers should focus on the most significant components and depict them clearly, as it was hard to explore all possible factors and their potential internal team interconnections.

According to Mashaly and Fernald [82], project managers encountered difficulties with SD modeling because, while subsystem integration may play a significant role in internal team modeling, most published studies only address and use two or three subsystems. Models of effectively integrated team management systems, according to Mashaly and Fernald [82], require careful consideration to design and cover several subsystems that may be essential for updating and refining the integrated modeling process as a whole. Some of the important modern challenges, such as team security, conflict, and retention, could be taken into account in the internal team resources modeling process to be incorporated evolutionary and holistically [82].

According to Manenzhe, Telukdarie, and Munsamy [29], researchers and practitioners had challenges when modeling SD due to continuing evolution. Most importantly, SD offered a temporary solution to complex project management challenges [29].

Small SD models helped decision-makers understand the conditions and causes of choice resistance, create immersive learning environments, overcome overconfidence, and promote the expansion of common knowledge across stakeholders [35]. As stated by Ghaffarzadegan, Lyneis, and Richardson [35], the SD profession ought to support policymakers in incorporating small SD models into their decision-making process on the internal team challenges.

Project managers had difficulties with SD modeling, according to Mashaly and Fernald [82], as a result of combining necessity: According to Mashaly and Fernald [82], the SD modeling approach was unable to fully characterize, analyze, and comprehend team behavior in several internal team difficulties.

Harms $et\ al.$ [31], reported that there were still unresolved issues with expert and computational programming of SD in project management. Bugalia, Maemura, and Ozawa [33] identified three major challenges regarding the internal team: changing the psychological framework, driving change, involving stakeholders, and effectively communicating and applying the strategy.

C. Advantages of Using SD

Mashaly and Fernald [82] asserted that Forrester utilized SD to simulate industrial dynamics for the first time in 1958. Since then, there has been a discernible improvement in the application of SD for planning, analysis, management, and comprehension. According to Eidin $et\ al.$ [80], project leaders should be able to conquer these difficulties and make sense of complicated phenomena by applying an SD methodology along with automated system models. According to David and Margaret [81], SD may be used to learn more about the makeup, behavior, and possible impacts of different circumstances and policies on intricate systems.

According to Eidin $et\ al.$ [80], SD is a cross-cutting concept that ought to be incorporated into the disciplines of scientific education as an encouraging means to deal with a range of societal issues. Mashaly and Fernald [82] suggested that employing SD to dissect a system into interconnected chains of stocks and flows that communicate with one another via both beneficial and detrimental feedback chains could be a helpful method for examining and evaluating a system's behaviors over time. The results of David and Margaret [81] indicate that the use of SD may be beneficial for evaluating theories on the origins and consequences of opportunities or systemic issues. Mashaly and Federald [82] assert that SD features made handling big, complicated, interconnected structures easier by saving time and effort—much like project management challenges in general.

Eidin $et\ al.$ [80] state that SD is a strategy that may be applied in a variety of sectors and disciplines to help with difficult problems and comprehend complicated phenomena. According to David and Margaret [81], SD can be used to assess different scenarios and approaches for system innovation or development and to disseminate the results and recommendations to decision-makers and interested parties. Mashaly and Federald [82] assert that the SD technique can be used to study and analyze the dynamic relationships that

take place inside systems as a whole. Mashaly and Federald [82] claim that the biggest benefit of the SD technique is its capacity to replicate a system's behavior and performance even in the absence of a prior evaluation and testing need.

According to David and Margaret [81], SD may be a useful tool for enhancing cognitive functions and transmitting knowledge. According to Mashaly and Federald [82], creating an SD has several benefits. Initially, by revealing the essential framework of the system, SD has the potential to streamline and enhance the clarity of intricate circumstances. Second, by advancing our understanding of the dynamics and architecture of complex systems, SD may make it possible to create extremely successful long-term transformation strategies. Furthermore, through the creation of well-established models and decision simulators, SD could support policy establishment and learning.

VII. Research Methodology of Internal Team Challenges

There were three main parts to the research methodology. The three crucial elements of the research strategy, are as follows:

i. Putting together a literature review. When assembling a literature review, keep the following points in mind:
   a. Investigate problems about internal team obstacles.
   b. Examine the rationale for project managers' typical lack of usage of SD modeling to resolve issues with internal teams.
   c. Put the finished literature review in the section designated by the literature review for this study.

ii. Utilizing five project management journal institutions to compare the literature review. Researchers examined sixty articles' worth of content from all of the publications to find out more
about the evaluation of internal team challenges. Finding the main reasons behind project managers' hesitation about using SD modeling to address problems with project internal team challenges was also essential. The author has gotten a total of 300 papers from five project management journal institutes (60 papers/journal multiplied by 5 journals). It is important to highlight that the 300 papers that were utilized to verify, validate, and corroborate the problem statement (internal team challenges) and literature review of the study were left out of the portion devoted to literature reviews. The five periodicals listed below were utilized:
   a. IGI Global journal. For information about the IGI Global journal, refer to [47].
   b. Project Management Research and Practice. For more information about the Project Management Research and Practice, refer to [48].
   c. South African Journal of Business Management. For more information about the South African Journal of Business Management, refer to [49]
   d. American Journal of Industrial and Business Management. For more information about the American Journal of Industrial and Business Management, refer to [50]
   e. Association for Project Management. For more information about the Association for Project Management journal, refer to [51].

iii. To validate and verify the study, the problem statement (internal team challenges) and literature review are the basis. Through the use of the Vensim software, the research offered a novel SD model. This new SD model has never been used to address the internal team challenges.

iv. testing and assessing the new SD model's output.

v. In the section titled "Results and Analysis," tabulate the findings.

VIII. Materials and Methods Used in Internal Team Challenges

The study made use of the following procedures and materials:

A. Study Design

The internal team difficulties that arise during project management provided the framework for the creation of this investigation. The subsequent research inquiries, which stem from issues faced by the team internally, also have an impact on the research design. This study's primary focus is on how SD can be used to address problems with internal team obstacles during the project management process. The secondary research questions are as follows:

i. Do project managers have issues with work assigned to internal teams?

ii. Are project managers knowledgeable about SD models, which can help with internal team problems during project management?

iii. Do project managers know about the issues that develop inside their teams?

iv. Can managers create and build SD models using mathematical formulas and programming skills to address issues with their teams?

v. Is an SD model that can address issues inside a team readily available for free download from the public domain (such as periodicals, little CDs, or Google)?

The authors conducted a descriptive phenomenology study using a literature review approach focused on internal team issues, utilizing open journals following the previously described study methodology.

B. Study Setting

The authors of the study collected sixty submissions (articles) to discuss internal team concerns from each of the five different journal institutions that were selected. Therefore, a total of 300 articles (5 multiplied by 60) were obtained for this investigation. The list of the five journal institutions is as follows:

i. IGI Global journal. For information about the IGI Global journal, refer to [47].

ii. Project Management Research and Practice. For details regarding the Project Management Research and Practice, refer to [48].
iii. South African Journal of Business Management. For information on the South African Journal of Business Management, refer to [49]
iv. American Journal of Industrial and Business Management. For details regarding the American Journal of Industrial and Business Management, refer to [50]
v. Association for Project Management. For more information about the Association for Project Management journal, refer to [51].

Regarding internal team challenges. These five journal institutions cover a variety of management-related topics, including system dynamics in management, engineering difficulties, project planning, business management challenges, and engineering management.

C. Sampling and Participant Recruitment

After study questions regarding internal team problems were submitted, the author was able to compile 300 papers and do research on each study question that was included in these publications. The 300 publications that were collected from the five journal institutions mentioned before produced insightful findings. The author's and articles' outputs about internal team challenges were constructed using an intentional criteria methodology for sampling in a joint study review about internal team challenges for each journal institution.

The journals were chosen with a variety of themes about issues faced by internal teams in mind. The author collected the data for this study by reading and analyzing 300 periodicals about internal team struggles. These 300 articles also address civil engineering management, leadership teams in businesses, management at hospitals, municipal governance, government agency management, managerial concerns, and industrial project management.

254 of the 300 articles provided thoughtful answers to the following question: Do project managers have issues with work assigned to internal teams? Refer to Table I and Figure 3. 254 problems with internal team obstacles in project management were discovered by the author. The aforementioned findings were included in all 254 publications from the five journal organizations listed above. Refer to Table I and Figure 3.

Out of 300 articles, 254 provided thoughtful answers to the following query: Are project managers knowledgeable about SD models, which can help with internal team problems during project management? Refer to Table II and Figure 4. Although project managers are aware of SD modeling, the author collected 278 publications that demonstrate that they do not apply it to internal team difficulties during project management. These conclusions were present in all 254 publications from the five journal organizations listed above. Refer to Table II and Figure 4.

Of the 300 articles, 243 offered a helpful response to the question: Do project managers know about the issues that develop inside their teams? Refer to Table III and Figure 5. The author found 243 articles that implied project managers are conscious of the problems posed by internal team issues during projects. These results were found in all 243 papers from the five journal organizations mentioned above. Refer to Table III and Figure 5.

Out of 300 publications, 243 papers addressed the following question satisfactorily: Can managers create and build SD models using mathematical formulas and programming skills to address issues with their teams? Refer to Table IV and Figure 6. The author discovered 243 articles with internal team issues that demonstrated managers' inability to design and develop SD using mathematical computations and programming. This results from the requirement for the programming and mathematical expertise required to produce SD simulations. These findings were included in each of the 243 papers that came from the five journal institutions listed above. Refer to Table IV and Figure 6.

Not a single one of the 300 publications adequately addressed the following question or offered a solution: Is an SD model that can address issues inside a team readily available for free download from the public domain (such as periodicals, little CDs, or Google)? Refer to Table V and Figure 7. The author discovered that there was not an SD model that could be downloaded and utilized to solve problems with internal team obstacles. These findings were included in each of the 300 papers that came from the five journal institutions listed above. Refer to Table V and Figure 7.

D. Data Collection

The 300-article study project came to an end when the degree of data redundancy regarding internal team challenges was reached, that was when no newly obtained or important information was added to later research. All of the data on problems with internal team challenges has been logged, and the study was complete.

E. Data analysis

The following data analysis of internal team issues that arise during project management will be conducted based on the answers to literature review research questions obtained from five journal institutions.

i. The author will conduct a statistical analysis to see if project managers struggle with assigning work to internal teams.
ii. The author will conduct a statistical analysis to assess project managers' knowledge of SD models, which can aid in resolving internal team issues during project management.
iii. The author will conduct a statistical analysis to assess project managers' awareness of team difficulties.
iv. The author will assess project managers' ability to design SD models utilizing mathematical formulas and programming skills to address team difficulties.
v. The author will do a statistical analysis to determine the number of SD models available for free download

from public domain sources (e.g., periodicals, CDs, or the internet) to address team difficulties.

TABLE I
RESULTS: DO PROJECT MANAGERS HAVE ISSUES WITH WORK ASSIGNED TO INTERNAL TEAMS?

Journal's Institution	The Number of Articles	Percentage of The Number of Articles Responding	1 Article Out of 60 as an Example
IGI Global journal (IGIGJ)	50	50/60 = 83.3333%	There is a yes answer to the question. Ref to [71]
Project Management Research and Practice (PMRP)	49	49/60 = 81.6666%	There is a yes answer to the question. Ref to [72]
South African Journal of Business Management (SAJBM)	51	51/60 = 85%	There is a yes answer to the question. Ref to [73]
American Journal of Industrial and Business Management (AJIBM)	57	57/60 = 95%	There is a yes answer to the question. Ref to [74]
Association for Project Management (APM)	47	47/60 = 78.3333%	There is a yes answer to the question. Ref to [75]

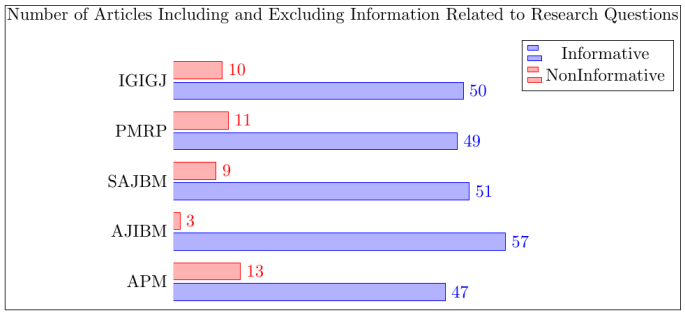


Fig. 3. Results: Do project managers have issues with work assigned to internal teams?

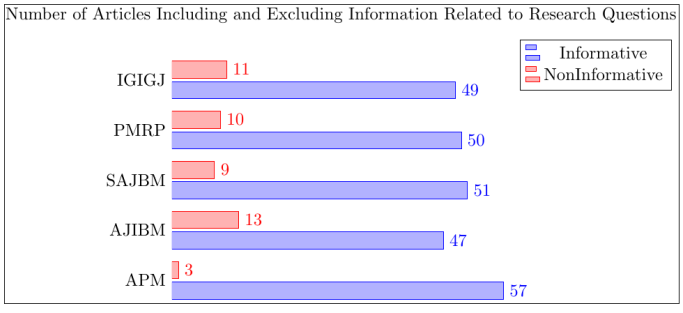


Fig. 4. Results: Are project managers conversant with SD models, which can resolve conflicts between teams?

TABLE II
RESULTS: ARE PROJECT MANAGERS CONVERSANT WITH SD MODELS, WHICH CAN RESOLVE CONFLICTS BETWEEN TEAMS?

Journal's Institution	The Number of Articles	Percentage of The Number of Articles Responding	1 Article Out of 60 as an Example
IGI Global journal (IGIGJ)	49	49/60 = 81.6666%	There is a yes answer to the question. Ref to [76]
Project Management Research and Practice (PMRP)	50	50/60 = 83.3333%	There is a yes answer to the question. Ref to [72]
South African Journal of Business Management (ASJBM)	51	51/60 = 85%	There is a yes answer to the question. Ref to [77]
American Journal of Industrial and Business Management (AJIBM)	47	47/60 = 78.3333%	There is a yes answer to the question. Ref to [78]
Association for Project Management (APM)	57	57/60 = 95%	There is a yes answer to the question. Ref to [79]

TABLE III
RESULTS: DO PROJECT MANAGERS KNOW ABOUT THE ISSUES THAT DEVELOP INSIDE THEIR TEAMS?

Journal's Institution	The Number of Articles	Percentage of The Number of Articles Responding	1 Article Out of 60 as an Example
IGI Global journal (IGIGJ)	48	48/60 = 80%	There is a yes answer to the question. Ref to [71]
Project Management Research and Practice (PMRP)	49	49/60 = 81.6666%	There is a yes answer to the question. Ref to [72]
South African Journal of Business Management (SAJBM)	47	47/60 = 78.3333%	There is a yes answer to the question. Ref to [73]
American Journal of Industrial and Business Management (AJIBM)	50	50/60 = 83.3333%	There is a yes answer to the question. Ref to [74]
Association for Project Management (APM)	49	49/60 = 81.6666%	There is a yes answer to the question. Ref to [75]

TABLE IV
RESULTS:
CAN MANAGERS PROGRAMMING SD MODELS USING MATHEMATICS TO SOLVE ISSUES WITH THEIR TEAMS?

Journal's Institution	**The Number of Articles**	**Percentage of The Number of Articles Responding**	**1 Article Out of 60 as an Example**
IGI Global journal (IGIGJ)	49	$49/60 = 81.6666\%$	Yes. indeed. Due to the lack of programming, design, and arithmetic knowledge, project managers are reluctant to apply SD improperly. Examine [76]. In addition, see Appendix A for additional details.
Project Management Research and Practice (PMRP)	47	$47/60 = 78.3333\%$	Yes. indeed. Due to the lack of programming, design, and arithmetic knowledge, project managers are reluctant to apply SD improperly. Examine [72]. In addition, see Appendix A for additional details.
South African Journal of Business Management (SAJBM)	50	$50/60 = 83.3333\%$	Yes. indeed. Due to the lack of programming, design, and arithmetic knowledge, project managers are reluctant to apply SD improperly. Examine [77]. In addition, see Appendix A for additional details.
American Journal of Industrial and Business Management (AJIBM)	49	$49/60 = 81.6666\%$	Yes. indeed. Due to the lack of programming, design, and arithmetic knowledge, project managers are reluctant to apply SD improperly. Examine [78]. In addition, see Appendix A for additional details.
Association for Project Management (APM)	48	$48/60 = 80\%$	Yes. indeed. Due to the lack of programming, design, and arithmetic knowledge, project managers are reluctant to apply SD improperly. Examine [79]. In addition, see Appendix A for additional details.

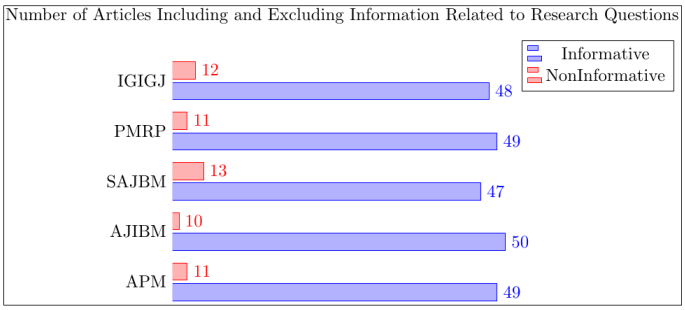


Fig. 5. Results: Do project managers know about the issues that develop inside their teams?

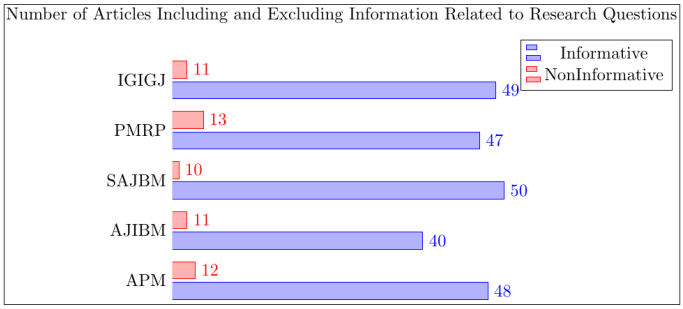


Fig. 6. Results: Can managers programming SD models using mathematics to solve issues with their teams?

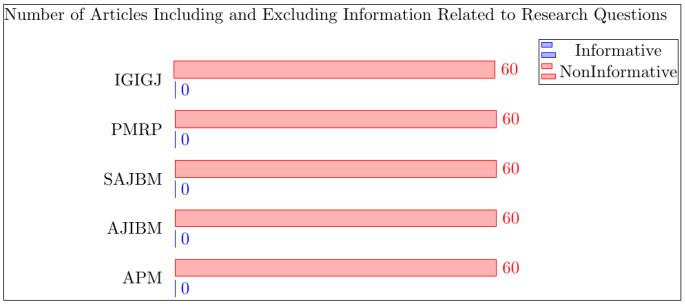


Fig. 7. Results: Is there an SD model to address issues inside a team readily available for free download?

IX. RESULTS AND DISCUSSION

Research in data analysis, experimentation, and literature review analysis served as the method's foundation. Moreover, theoretical conclusions drawn from the literature review were contrasted with the outcomes of the experiments. A reference to Table VI is made. Following analysis of the study questions, the author was able to gather 300 papers, using them to investigate subjects connected to internal team challenges for each research question. From 300 papers gathered from the five journal organizations previously stated, practical solutions were discovered. Using a deliberate criterion sampling technique, the author's and articles' outputs were produced through the joint research review procedure inside each journal institution.

The journals have been chosen to address various kinds of engineering project management internal team concerns. The

TABLE V
RESULTS: IS THERE AN SD MODEL TO ADDRESS ISSUES INSIDE A TEAM READILY AVAILABLE FOR FREE DOWNLOAD?

Journal's Institution	**The Number of Articles**	**Percentage of The Number of Articles Responding**	**1 Article Out of 60 as an Example**
IGI Global journal (IGIGJ)	0	$0/60 = 0\%$	Not even one SD model should be downloaded from this publication and used immediately.
Project Management Research and Practice (PMRP)	0	$47/60 = 78.3333\%$	Not even one SD model should be downloaded from this publication and used immediately.
South African Journal of Business Management (SAJBM)	0	$50/60 = 83.3333\%$	Not even one SD model should be downloaded from this publication and used immediately.
American Journal of Industrial and Business Management (AJIBM)	0	$49/60 = 81.6666\%$	Not even one SD model should be downloaded from this publication and used immediately.
Association for Project Management (APM)	0	$48/60 = 80\%$	Not even one SD model should be downloaded from this publication and used immediately.

TABLE VI
COMPARING THE EXPERIMENTAL FINDINGS OF THE STUDY WITH THE LITERATURE REVIEW

Theoretical Perspectives from the Literature Review	**Outcomes of the Experiments Run for This Research**
Figure 1 and Figure 2 were found to be available as the traditional method to solve internal team challenges for reference to support claims [8] [82], hence no SD was utilized to address the issue.	The authors were motivated to tackle the problem with SD.
Due to SD modeling's complexity, most project managers dislike it, especially those who have never worked in IT, information theory, control elements, engineering, programming, feedback loops, or mathematical skills [28], [30], [34]. Additionally, mathematical principles programming, stock analyzing, fault dealing, prediction models, inputs from cognitive science, causal loop schematics, models, and latencies are all required for SD modeling [29], [31], [35], [84]. Refer to Figures 8 and 9 for a summary explanation.	In the present article, authors have established the SD model to handle the difficulties of internal team challenges. In the beginning, the authors constructed the first SD model with algebraic formulas and units, subsequently followed by conducting simulations using the Vensim application as illustrated in Figure 8 and Figure 9, respectively. The end product of the simulation demonstrated that all formulas and units allocated to the model are interacting with other components as intended. Furthermore, the results claimed that the flows and control components depend on each other to function.

author investigated and analyzed 300 articles before obtaining the data for the current research. These 300 articles discuss a variety of themes, including development in the disciplines of engineering, hospitals, towns, leadership teams in businesses, office administration concerns, departmental management, and commercial project management.

A. Literature Review Research with Data Analytic Questions

This study's crux primary research questions are:

i. Do project managers have issues with work assigned to internal teams? The question has a yes response. Internal team difficulties are challenging to project managers, as reported by 78.3333% to 95% of respondents. Refer to Table I and Figure 3.
ii. Are project managers knowledgeable about SD models, which can help with internal team problems during project management? There is a yes response to the query. It was validated by 78.3333% to 95.3% that SD can address issues with internal team challenges. Refer to Table II and Figure 4.
iii. Do project managers know about the issues that develop inside their teams? There is a yes response to the query. Confirmation of the awareness of the issues faced by internal teams ranged from 78.3333% to 83.3333%. Please consult Figure 5 and Table III.
iv. Can managers create and build SD models using mathematical formulas and programming skills to address issues with their teams? Yes, It is. The 78.3333% to 83.3333% results showed that managers could use SD to programmatically handle mathematically-based internal team difficulties while maintaining ethical conduct. Please refer to Table IV and Figure 6. Project managers are afraid to implement SD incorrectly because they lack

the necessary programming, design, and math skills.

v. Is an SD model that can address issues inside a team readily available for free download from the public domain (such as periodicals, little CDs, or Google)? Not a ready-to-use SD model is available from these five publications for download. Please refer to Figure 7 and Table V.

B. Experimental Findings Using the Suggested SD Model

The SD model was created by the authors of this study to address internal team issues. The initial SD model was created by the authors using algebraic equations and components, and it was then simulated using the Vensim program, as shown in Figures 8 and 9, respectively.

The simulation's outcomes demonstrated that every equation and unit included in the model is interacting with every other unit as it should. The outcomes also showed that the control components and flows are interdependent. The following reasons support the aforementioned statement:

i. Figure 10 shows that there was a dependency of $Individual$ (or team) and $Tasks/Month$ (or sketch representation taken from internal team challenges) when comparing $NonConflicted\ Team$ and $work\ flow$ as SD models.

ii. Figure 11 and Equation 1 show that the changes in $improvements\ recognition\ rate$ depend on $Unnoted\ Improvements$ and $conflicts\ detection\ time$.

$$improvements\ recognition\ rate = Unnoted\ Improvements\ /conflicts\ detection\ time \quad (1)$$

Units: Tasks/Month

iii. Equation 2 and Figure 12 indicate that $improvements\ recognition\ rate$ and $work\ flow$ are dependent on changes in $Tasks\ Ignored$.

$$Tasks\ Ignored = INTEG(improvements\ recognition\ rate - work\ flow, first\ task\ explanation) \quad (2)$$

Units: Tasks

iv. Figure 13 and Equation 3 show that the changes in $Finished\ Tasks$ rely on $improvements\ recognition\ rate$ and $work\ flow$.

$$Finished\ Tasks = INTEG(work\ flow - improvements\ recognition\ rate,\ 0) \quad (3)$$

Units: Tasks

v. Figure 14 and Equation 4 show that the changes in $Unnoted\ Improvements$ rely on $work\ flow$, $competency$, and $improvements\ recognition\ rate$.

$$Unnoted\ Improvements = INTEG(work\ flow * (1 - competency) - improvements\ recognition\ rate,\ 0) \quad (4)$$

Units: Tasks

v. Equation 5 also provides dependencies on other variables.

$$NonConflicted\ Team = INTEG(\ briefing\ completions - dismissals * ZIDZ(NonConflicted\ Team, total\ workforce), 0) \quad (5)$$

Units: Individual

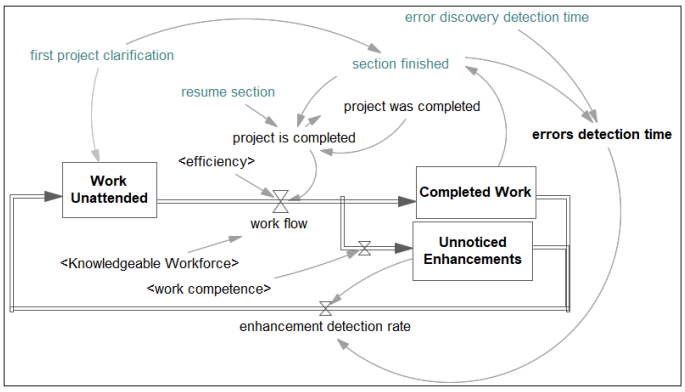


Fig. 8. Project Managers Designed the SD Model before it was Ran

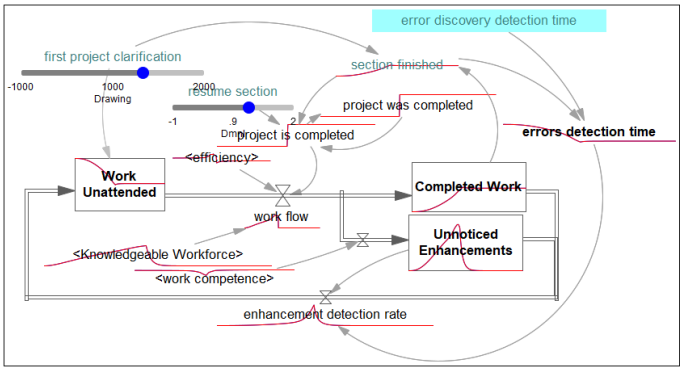


Fig. 9. An SD Model Designed with Project Managers in Consideration As it was Running

Most project managers hate SD modeling because of its complexity, especially if they don't have any background in programming, literature review analysis, or information technology. SD modeling requires the use of mathematical programming, engineering, logical decision-making, causal loop diagrams, statistical models, feedback loops, delays, control elements, stock understanding, defect handling, and computing models. It also requires feedback on data theory.

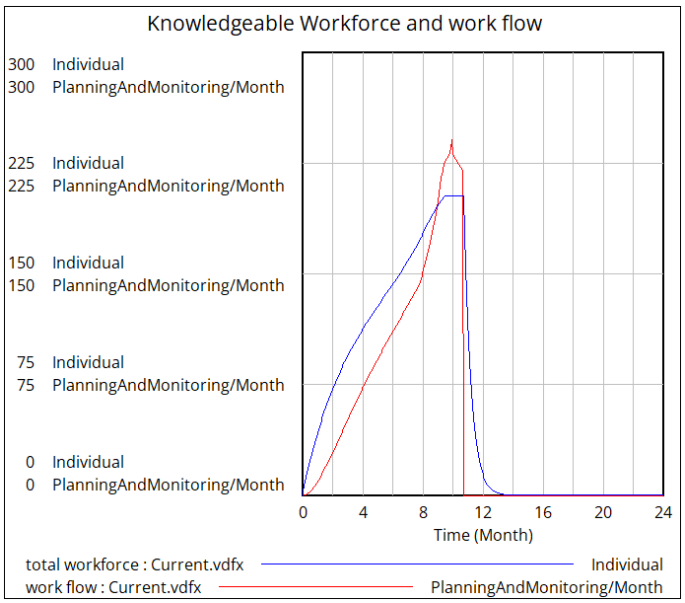


Fig. 10. Comparing Work Flow and NonConflicted Teams

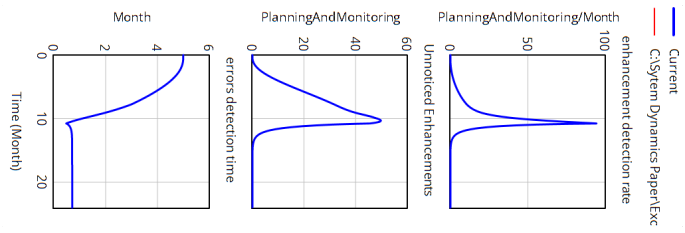


Fig. 11. Causes of Improvements Recognition Rate

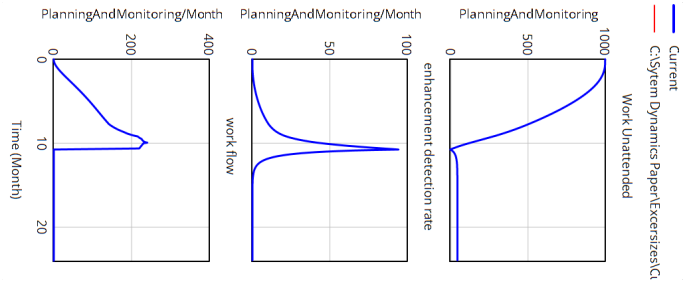


Fig. 12. Causes of Tasks Ignored

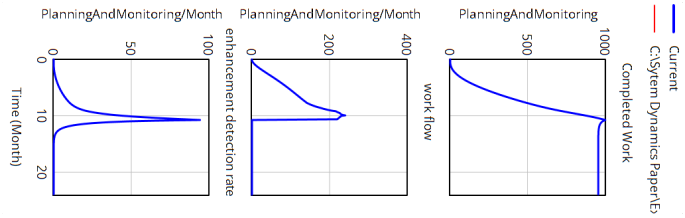


Fig. 13. Causes of Finished Tasks

The SD model was put forth by the paper's author as a tool to help managers with project management.

Based on the results of the experiment, the authors developed an SD model that project managers can use to address problems with scheduling and monitoring. The SD model illustrated how flows and control systems might be modified following the number of individuals and projects in need of scheduling and oversight. When it comes to resolving planning and transportation-related problems, project managers who are not proficient in programming or mathematics might find the newly suggested SD model useful.

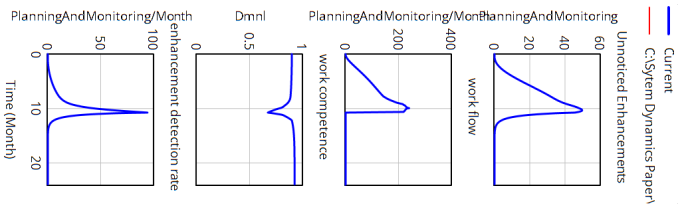


Fig. 14. Causes of Unnoticed Improvements

X. Conclusion, Future Work and Limitation of the Study

The majority of project managers hate SD modeling because of its complexity, especially if they don't have any background in coding, engineering, mathematical computation, IT, logic-based decision-making, software engineering, causal loop diagrams, computational models, information feedback principle, flow using, control elements, programming languages, or delays. The authors put forth the SD model as a tool to help managers with internal team challenges.

The study used literature review and experimental methodologies to look into internal team issues. The results of the literature review research show that between 78.3333% to 95% of respondents thought that managing an internal team was challenging for project managers. 78.3333% to 95.3% of respondents confirmed that SD can handle problems with the internal team. A range of 78.3333% to 83.3333% showed that managers might use SD to use mathematics to address internal team problems while continuing to act responsibly, however, it is impossible to develop an SD model without programming and mathematical analytics expertise. These five publications do not offer a plug-and-play SD model for download. Through experimental study, the authors built and coded a novel SD model that might help managers resolve internal team challenges.

A. Future Work

The authors propose to use the SD model in a future study to address planning, monitoring, scope, and cost-estimating difficulties.

B. Limitation of the Study

The readers of this article should note that while the study offered a new SD model to address issues with internal teams, it does not address issues with scope, cost estimation, or cohesiveness, among other project management issues. The various SD models are required for different problems.

REFERENCES

[1] J. Aceron, "Top 3 Project Management Challenges And How To Overcome Them, Nimble Humanize Work, pp. 1-17, 2023.

[2] Fastercapital "Project scope: Project Scope and Its Impact on Accurate Cost Estimation,Fastercapital, pp. 1-299, 2024.

[3] D. Garg, "Characterizing Project Scope Attributes and their Influence on the Software Estimation Process," Massachusetts Institute of Technology, pp. 1-72, 2023.

[4] S. S. N. Masombuka and X. C. Thani, "Challenges and Successes of the Government-wide Monitoring and Evaluation System," Sabinet African Journals, pp. 1-18, 2023.

[5] Department of Planning, Monitoring, and Evaluation, " Revised Annual Performance Plan 2022/2023," Department of Planning, Monitoring, and Evaluation, pp. 1-136, 2023.

[6] S. Chellapa, "5 Teamwork Challenges Every Team Encounters!," Engagedly, pp. 1-5, 2023.

[7] G. G. Benarkuu and E. Katere, The Effects of Dynamic Team-Building Intervention on Internal Communication in the Hospitality Industry in Sunyani Municipality, Ghana," World Journal of Advanced Research and Reviews, pp. 1-9, 2023.

[8] Faster Capital "Project Scope: Project Scope and Its Impact on Accurate Cost Estimation," Faster Capital, pp. 1-89, 2024.

[9] B. Sushmith, "A Brief Guide to Conflict Management Approaches," Sprintzeal, pp. 1-10, 2023.

[10] T. Althiyabi and M. R. J. Qureshi, "Predefined Project Scope Changes and its Causes for Project Success," International Journal of Software Engineering and Applications vol. 1 2, no: 2/3, pp: 45-56, 2021.

[11] EvalCommunity, "What are the principles of monitoring and evaluation?," EvalCommunity, pp.1-8, 2023.

[12] Team Asana, "What is the project management triangle and how can it help your team?," Team Asana, pp.1-7, 2024.

[13] Indeed Editorial Team, "12 Project Management Challenges and How To Solve Them," Indeed Editorial Team, pp. 1-8, 2023.

[14] B. Amade and E. O. P. Akpan, "Project Cost Estimation: Issues and the Possible Solutions," International Journal of Engineering and Technical Research (IJETR), vol: 2, no: 5, 2014.

[15] A. Monnappa, "Project Scope Management and Its Importance in 2024," Simplilearn, pp. 1-23, 2023.

[16] A. Rodrigues, The Role of System Dynamics in Project Management: A Comparative With Traditional Models, Internationa System Dynamics Conference, pp. 1-24, 1994.

[17] C. Tristancho, "Gantt Chart vs. Kanban Board: Pros, Cons, Similarities & Differences," ProjectManger.com, pp. 1-34, 2024.

[18] E. Chappell, "Gantt Chart vs. Timeline: What Are They and How to Use Them," Clickup. pp. 1-15, 2023.

[19] J. Stermann, "System Dynamics Modelling for Project Management," Themys.sid.uncu.edu.ar, pp. 1-12, 1992.

[20] K. S. Sima, "Executing, Monitoring and Controlling a Project, the Right Way," Researchgate, pp. 1-52, 2022.

[21] Educational Foundation of Project Management Institute, "Project Management Guide 2022/23," Educational Foundation of Project Management Institute, pp. 1-26, 2023.

[22] K. Eby, "Guide to Project Monitoring and Control Phase of Project Management,"SmartSheet.com, pp 1-16, 2022.

[23] H. I. Ibrahim and S. K. Silas, "Planning for Monitoring and Evaluation on Performance of Girl's Education Projects of Public Primary School in Baidoa, Somalia," International Journal of Scientific and Research Publications, vol. 13, no. 10, pp. 1-8, 2023.

[24] G. G. Benarkuu and E. Katere, "The Effects of Dynamic Team-Building Intervention on Internal Communication in the Hospitality Industry in Sunyani Municipality, Ghana," World Journal of Advanced Research and Reviews, vol. 18, no: 03, pp. 1567–1575, 2023.

[25] A. Htet, S. R. Liana, T. Aung, and A. Bhaumik, "Engineering Management: A Comprehensive Review of Challenges, Trends, and Best Practices," Journal of Engineering and Pedagogy, vol. 1, no: 1, pp. 1-7, 2023.

[26] J. Birt, "14 Teamwork Challenges and How To Overcome Them," Indeed, pp. 1-20, 2023.

[27] M. A. Kazemitabar, S. P. Lajoie, and T. Li, "A Classification of Challenges Encountered in Complex Teamwork Settings," International Journal of Computer-Supported Collaborative Learning, vol. 17, no: 1, 2022.

[28] D. Rumeser and M. Emsley, "Key Challenges of System Dynamics Implementation in Project Management," 3rd International Conference on New Challenges in Management and Organization," pp. 22-30, 2016.

[29] M. T. Manenzhe, A. Telukdarie, and M. Munsamy, "Maintenance Work Management Process Model: Incorporating System Dynamics and 4IR Technologies," Emerald, pp. 1-32, 2022.

[30] S. E. Amin, H. M. Abdul-kader, and A. H. Elsaid "Using Systems Dynamics in Modeling of Dynamic Capabilities: A Review Study," International Journal for Computers and Information. vol. 10, no. 3, 2023.

[31] J. Z. Harms, J. J. Malard-Adam, J. F. Adamowski, A. Sharma, and A. Nkwasa, "Dynamically Coupling System Dynamics and SWAT+ Models using Tinamït: Application of Modular Tools for Coupled Human–Water System Models," HESS, vol. 27, pp. 1683–1693, 2023.

[32] D.J. Crookes $et\ al.$, "System Dynamic Modelling to Assess Economic Viability and Risk Trade-Offs for Ecological Restoration in South Africa," Repository.up.ac.za, pp. 1-29, 2023.

[33] N. Bugalia, Y. Maemura, and K. Ozawa, "A System Dynamics Model for Near-Miss Reporting in Complex Systems" Safety Science, vol.142, pp. 1-7, 2021.

[34] F. B. B. Nasution and N. E. N. Bazin, "Public Policymaking Framework Based on System Dynamics and Big Data," International Journal of System Dynamics Applications, vol. 7, no. 4, pp. 38-53, 2018.

[35] N. Ghaffarzadegan, J. Lyneis, and G. P. Richardson, "How Small System Dynamics Models Can Help the Public Policy Process," Albany, pp. 1-38, 2023.

[36] https://epojournal.net/

[37] https://hbr.org/subscriptions?utm_medium=paidsearch&utm_source=google&utm_campaign=subscribetohbr_gbb_intl&utm_term=Brand&tpcc=paidsearch.google.brand&gad_source=1&gclid=Cj0KCQjw0MexBhD3ARIsAEI3WHIHLS7arAb1LXk QjluK2t0hduzlTX27a4LVaWbvltVdYCCcBUbvNxAaAvqgEALw_wcB

[38] https://www.tandfonline.com/journals/tjcm20

[39] https://www.emeraldgrouppublishing.com/journal/ijmpb

[40] https://shop.elsevier.com/journals/journal-of-engineering-and-technology-management/0923-4748

[41] https://www.sciencedirect.com/journal/international-journal-of-project-management

[42] http://www.ppml.url.tw/EPPM_Journal/

[43] https://www.projectmanagement.com/pm-network/#_=_

[44] https://www.picmet.org/main/

[45] https://www.sciencedirect.com/journal/project-leadership-and-society

[46] https://www.forbes.com/forbes-magazine/?sh=2514124730a4

[47] https://www.igi-global.com/journal/international-journal-information-technology-project/1103

[48] https://epress.lib.uts.edu.au/journals/index.php/PMRP/index

[49] https://sajbm.org/index.php/sajbm

[50] https://www.scirp.org/journal/ajibm/

[51] https://www.apm.org.uk/resources/

[52] E. Bingham, G. E. Gibson Jr, M. E Asmar, "Best Practices in Preconstruction Services for Transportation Projects," Engineering Project Organization Journal (EPOJ), p8, 1-12, 2018.

[53] HBR Editors, "The Four Phases of Project Management," HBR, pp. 1-10, 2016.

[54] https://novapublishers.com/shop/volume-13-issue-2-international-journal-of-construction-project-management/

[55] https://discovery.ucl.ac.uk/id/eprint/10101944/1/Attached

[56] https://www.sciencedirect.com/science/article/abs/pii/S0923474810000238

[57] https://img1.wsimg.com/blobby/go/d0dd54db-2225-42e2-a2f9-e42e4b1f907b/downloads/EPOJ

[58] https://store.hbr.org/product/system-dynamics-modeling-tools-for-learning-in-a-complex-world/CMR205

[59] https://www.tandfonline.com/doi/full/10.1080/15623599.2020.1854930

[60] https://www.sciencedirect.com/science/article/abs/pii/S0923474803000183

[61] https://ajpojournals.org/product/international-journal-entrepreneurship-project-management/

[62] http://www.ppml.url.tw/EPPM_Journal/volumns/13_02_May_2023/ID_526.htm

[63] https://www.projectmanagement.com/articles/391591/damage-control–project-management-in-the-midst-of-a-crisis

[64] https://www.picmet.org/new/Conferences/23/bulletin_23.pdf

[65] A. Hallstrom, P. Bosch-Sijtsema, ""I Can Say Things I Wouldn't Normally Say": Changing Project Delivery Implementation and Social Networks as Drivers of Institutional Change in Nordic Infrastructure Projects," pp. 1-13, 2024.

[66] R. J. Chapman, "The Role of System Dynamics in Understanding the Impact of Changes to Key Project Personnel on Design Production within Construction Projects," International Journal of Project Management, vol. 16, no. 4, pp. 235-247, 1998.
[67] T. T. Akano *et al.*, "Process Optimization of Engineering Work Request in an Offshore Environment," Journal of Engineering, Project, and Production Management, vol. 13, no. 1, 20-29, 2023.
[68] M. Gray and A. Shahidi, "Applying The Principles of System Dynamics in Project Risk Management," Project Management Institute, pp. 1-15, 2011.
[69] G. Schuh, M. Engel, T. Drescher, and K. Apfel, "Comprehensive Technology Exploitation Using System Dynamics," PICMET, pp. 1-10, 2014.
[70] R. Levitt, J. Pollack, and J. Whyte, "Leadership and the Dynamics of Projects: Ray Levitt's Insights on Project Leadership," Project Leadership and Society, pp. 1-15, 2024.
[71] E. Cano and G. Ion, "Curriculum Development through Competency-Based Approach in Higher Education," pp. 1-17, 2014. Available: https://www.igi-global.com/dictionary/curriculum-development-through-competency-based/28124
[72] UTC ePress, "Project Management Research and Practice," Project Management Research and Practice, pp. 1-234, 2018. Available: https://epress.lib.uts.edu.au/journals/index.php/PMRP/issue/view/455
[73] R. L. Thokoa, V. Nadioo, and T. Herbst, "An Exploration of Internal Branding at the National Treasury of South Africa," South African Journal of Business Management, vol. 53, no. 1, pp. 1-19, 2022. Available: https://sajbm.org/index.php/sajbm/article/view/2593
[74] Y Huang, J. Ye, and Z. Gao, "Study on Team Stability Based on the Perspective of Knowledge Potential," American Journal of Industrial and Business Management, pp. 1-30, 2022. Available: https://www.scirp.org/journal/ajibm/?utm_campaign=826331897_110518896423&utm_source=lixiaofang&utm_medium=adwords&utm_content=dsa-786034580569&gad_source=1&gclid=CjwKCAjw3NyxBhBmEiwAyofDYUvmF-qjWe-aVXBemw2Zao9sBkt1yyliNbjbSwhKUeesFSNyUVZjBoC4gQQAvD_BwE
[75] D. Waller, "Five Ways to Keep your Project Team Motivated," Association for Project Management, pp. 1-6, 2024. Available: https://www.apm.org.uk/blog/five-ways-to-keep-your-project-team-motivated/
[76] A. T. Azar, "International Journal of System Dynamics Applications," IGI Global Journal, pp. 1-105, 2012. Available: https://www.igi-global.com/journal/international-journal-system-dynamics-applications/51803
[77] S. Khumalo and T. Du Plessis, "Commercialisation Dynamics System Principles and Support Units of Entrepreneurial Universities," South African Journal of Information Management, vol. 26, no. 1, pp. 1-17, 2024.
[78] Gintautas P. Kamuntavičius, and G. Kamuntavičius, "Relativistic Dynamics of a Quantum System," Journal of Applied Mathematics and Physics, vol.12, no.4, April 30, 2024. Available: https://www.scirp.org/journal/journalarticles?journalid=2436
[79] N. Fewings, "Team Lead Succeed – Helping You and Your Team Achieve High-Performance Teamwork," APM, pp. 1-23, 2024. Available: https://www.apm.org.uk/news/team-lead-succeed-helping-you-and-your-team-achieve-high-performance-teamwork-2/
[80] E. Eidin, T. Bielik, and I. Touitou, *et al.*, "Thinking in Terms of Change over Time: Opportunities and Challenges of Using System Dynamics Models," J Sci Educ Technol vol. 33, pp. 1–28, 2024. Available: https://doi.org/10.1007/s10956-023-10047-y
[81] R. David and E. Margaret, "Key Challenges of System Dynamics Implementation in Project Management," Procedia - Social and Behavioral Sciences 230, pp. 22–30, 2016.
[82] A. F. Mashaly and A. G. Fernald, "Identifying Capabilities and Potentials of System Dynamics in Hydrology and Water Resources as a Promising Modeling Approach for Water Management,"MDPI, pp. 1-24, 2020.
[83] M. S. Sabiroh, M. Sharifah, and J. Nurul, "Exploring the Conflict Management Process: A Case Study of the Department of Labour in Malaysia," Jurnal Intelek, vol. 16, no. 1, pp. 7-16, 2021.
[84] T. D. Phan, E. Bertone, and R. A. Stewart, "Critical Review of System Dynamics Modelling Applications for Water Resources Planning and Management," ResearchGate, vol. 2, no. 3, 2021.

[**Dr. Khumbelo Difference Muthavhine** is a Ph.D. graduate from the University of South Africa and a follower of IEEE and MDPI journal institutions. He was born in Ha-Vhangani village in Venda, in the Limpopo province of South Africa. He has a BSc degree in mathematics and physics from the University of Venda. He obtained a BSc conversion in Electrical Engineering from the University of Cape Town and conducted his practicals at the University of Stellenbosch. He obtained a BTech degree in electrical engineering from Tshwane University of Technology. He obtained an M.Tech. and a PhD degree from the University of South Africa. He has published many papers with IEEE, MDPI, Academia, and the 2018 International Conference on Information and Communications Technology (ICOIACT) proceedings. His specialties are cryptography, network security, and engineering management.

A Deep Learning Framework for Heartbeat Classification Using ECG and MFCC Features

Remya Madhavan, Swati Singh, Vishal Dhanda
Department of Electrical and Electronics Engineering
NITTE Meenakshi Institute of Technology, Bengaluru, Karnataka
{remya.madhavan, 1nt23ee052.swati, 1nt23ad060.vishal}@nmit.ac.in

***Abstract*—Our research addresses early identification of cardiac irregularities is vital to improving patient outcomes and reducing long-term cardiovascular risk. This paper presents a lightweight Convolutional Neural Network (CNN) framework that leverages both time-domain ECG features and Mel-Frequency Cepstral Coefficients (MFCCs) derived from heart sound signals. The system achieves efficient binary classification of heartbeats (normal vs abnormal) and demonstrates strong potential for integration into real-time diagnostic tools. The proposed hybrid architecture is trained and evaluated on clinically diverse datasets and maintains accuracy with fewer than 100,000 parameters, enabling deployment in portable systems.The hybrid model integrates features from both ECG and PCG data, where ECG provides electrical signal patterns and MFCCs extracted from PCG represent acoustic characteristics. This combination takes advantage of ECG's timing accuracy and the frequency-based information from heart sounds to improve classification performance and ensure reliability across various clinical settings.**

***Index Terms*—ECG, MFCC, Phonocardiogram, CNN, Heartbeat Classification, Signal Processing, Deep Learning**

I. Introduction

Cardiovascular diseases (CVDs) are the leading cause of death globally, necessitating the development of early and accessible diagnostic systems. Heartbeat abnormalities are a kind of cardio- vascular condition. If these disorders are not detected in their early stages, it is possible for blood to congeal within the blood arteries, which can lead to heart failure, as well as other potentially deadly diseases. Therefore, using different signal processing methods to analyze and classify heart sound signals is an effective way to find cardiac problems [3]-[4]. Conventional auscultation relies on a physician's auditory skills to interpret heart sounds; however, this subjective approach can miss subtle abnormalities, particularly in noisy or resource-limited environments.[1]

Recent advances in signal processing and machine learning have enabled the transformation of raw biosignals—such as electrocardiograms (ECGs) and phonocardiograms (PCGs)—into discriminative features suitable for automated analysis[1]. This work introduces a CNN-based system that integrates both ECG morphology and MFCC representations of PCG signals for accurate and real-time heartbeat classification.[2]

II. Data Collection and Preprocessing

The dataset used in this study contains synchronized ECG and heart sound recordings acquired from both clinical and field conditions. Each recording includes annotations indicating cardiac events like S1(The first heart sound is generated when the mitral and tricuspid valves close. This sound signifies the start of the systolic phase, during which the heart muscles contract to push blood out. Compared to the second heart sound, S1 tends to have a lower pitch and a longer duration), S2 (The second heart sound arises from the closure of the aortic and pulmonary valves. It marks the end of systole and the onset of diastole, the phase when the heart relaxes and chambers refill with blood. S2 is generally shorter in length and has a higher frequency than S1)[3], systole, and diastole.

ECG signals were filtered using a 4th-order zero-phase Butterworth band-pass filter (20–400 Hz) to suppress baseline wander and high-frequency noise. Heart sound signals were windowed using 25 ms Hamming windows with a 10 ms stride to prepare for MFCC extraction.

Recordings were segmented into overlapping frames (25ms window, 10ms hop). For each frame: 1) Compute a short-time Fourier transform with nfft = 512. 2) Apply a triangular filter bank of 26 mel bands. 3) Derive 13 MFCCs via discrete cosine transform. The resulting sequence of MFCC vectors (size 13 × T) forms the input to our CNN. No additional filtering or augmentation was applied

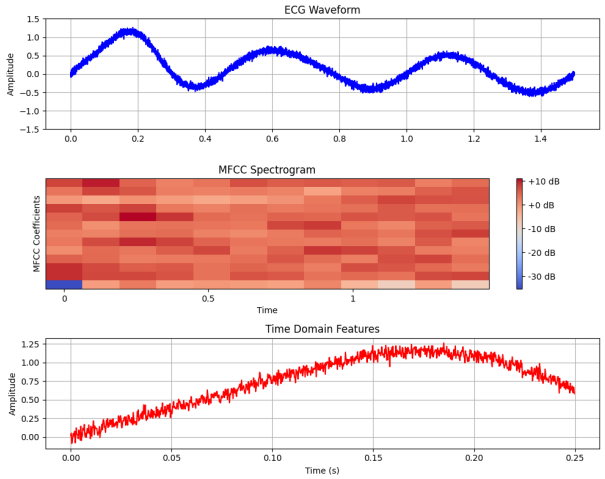


Fig. 1. Top: Denoised ECG waveform. Middle: MFCC representation of heart sound. Bottom: Time-domain envelope of PCG signal. Each layer contributes distinct diagnostic features.

III. Model Architecture

The proposed convolutional neural network (CNN) is specifically constructed to classify heart sound signals using Mel-Frequency Cepstral Coefficients (MFCCs) derived from phonocardiogram (PCG) recordings. The input to the model is a feature matrix with dimensions $13 \times 128 \times 1$, representing 13 MFCC bands across 128 temporal frames.

The network initiates with a 2D convolutional layer consisting of 32 filters, each of size 3×3. This layer captures localized patterns from the input and produces an output tensor of shape $(11, 126, 32)$. A batch normalization layer is immediately applied to standardize activations, thereby improving learning stability and convergence rate.

Subsequently, a 2×2 max pooling layer is used to downsample the feature map, reducing its spatial resolution to $(5, 63, 32)$. A second convolutional layer with 64 filters further processes the data, yielding an output of shape $(3, 61, 64)$. Another batch normalization layer is added to regularize and support the learning process.

A second 2×2 max pooling operation is then performed, compressing the feature map to $(1, 30, 64)$. This compact representation is flattened into a one-dimensional vector of size 1920, which serves as input to the fully connected layers.

The dense layer comprises 128 neurons activated by the ReLU function. To prevent overfitting, a dropout layer with a dropout rate of 0.5 is employed, randomly deactivating 50% of the neurons during training.

The output layer is a dense layer with 4 neurons, each representing a distinct heartbeat class. During inference, the softmax activation function is applied to interpret these outputs as class probabilities.

The total number of trainable parameters is maintained below 100,000, ensuring the model's suitability for deployment on low-resource devices such as mobile platforms or embedded medical systems. This architecture effectively balances computational efficiency and diagnostic accuracy by extracting discriminative features from acoustic cardiac signals.

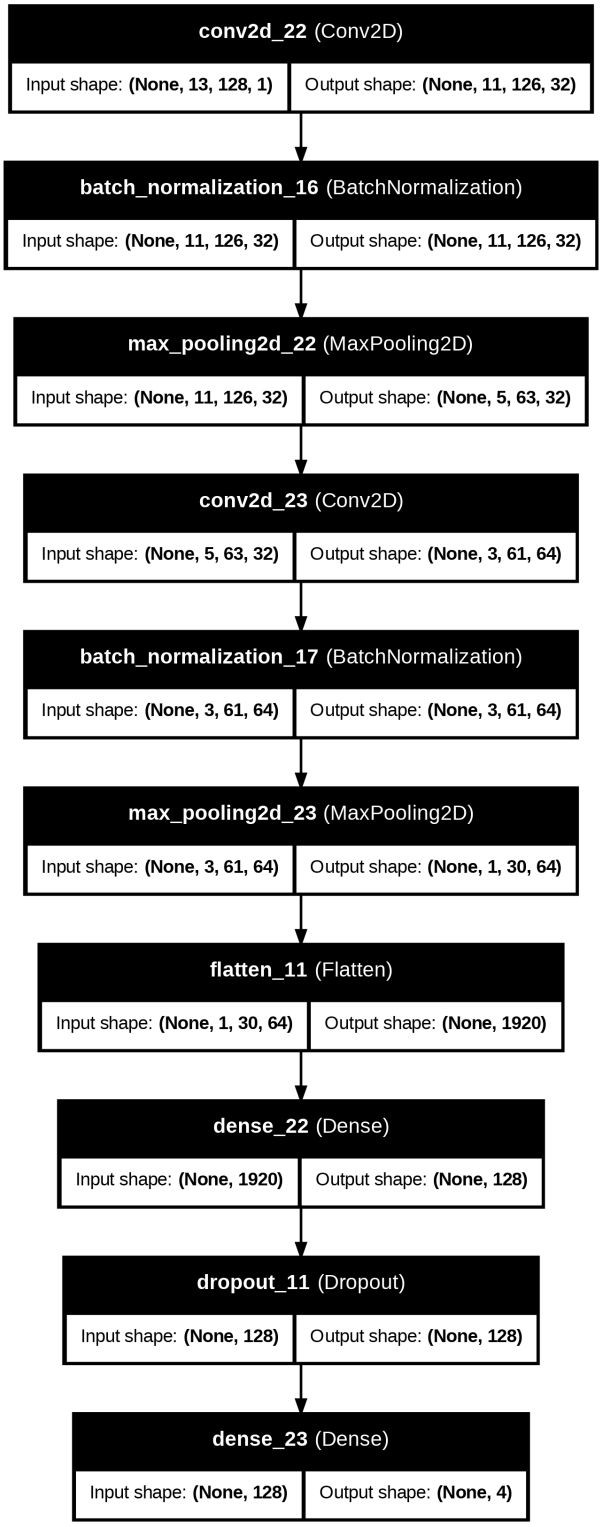


Fig. 2. CNN architecture for heartbeat classification. Compact yet expressive, it processes MFCC maps and ECG-derived matrices with minimal computational load.

IV. MFCC Feature Extraction

MFCCs are derived by applying short-time Fourier transform (STFT) followed by Mel filter bank integration and discrete

cosine transformation. These coefficients mimic human auditory perception, focusing on frequency ranges most relevant to cardiac acoustics. Each PCG signal is converted into a 2D feature matrix of size $13 \times T$, where 13 MFCCs are extracted per frame.

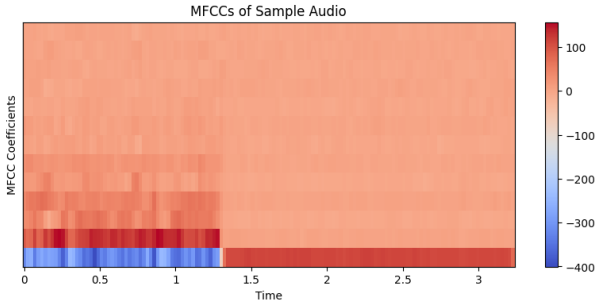


Fig. 3. MFCC spectrogram of a heart sound segment. Peaks in the lower bands often correspond to key cardiac cycles such as systole and diastole.

V. Evaluation and Results

The model was trained using the Adam optimizer with a learning rate of 1×10^{-4} for 20 epochs. Early stopping was employed to prevent overfitting. Evaluation was conducted on a separate test set using accuracy, precision, recall, F1-score, and a confusion matrix.

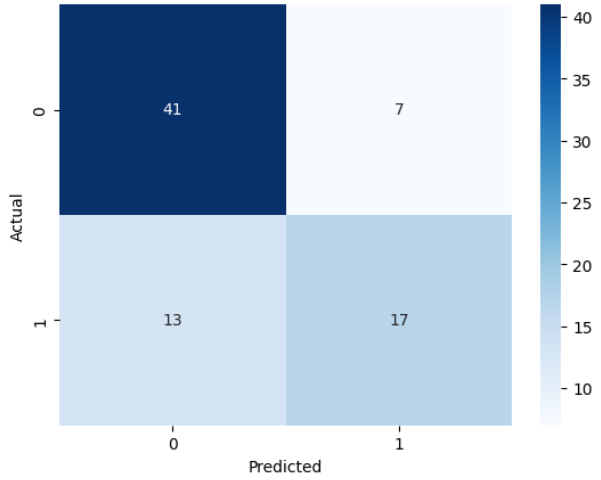


Fig. 4. Confusion matrix on test data showing high true positive and true negative rates. Misclassifications are primarily minor-type abnormalities.

TABLE I: Classification Report of the Proposed CNN Model

Class	Precision	Recall	F1-score	Support
0	0.76	0.85	0.80	48
1	0.74	0.74	0.74	78
2	0.71	0.57	0.63	30
Accuracy	0.74			
Macro Avg	0.73	0.71	0.72	78
Weighted Avg	0.74	0.74	0.74	78

The model achieved 74.3% accuracy, 56.6% sensitivity, 85.4% specificity, and an AUC of 71. Its compact nature (under 100K parameters) allows for real-time inference even on low-power edge devices. The source code for the heartbeat classification project is publicly available at:

https://github.com/Swati-cod/heartbeatclassification.git

VI. Conclusion

This study proposes a hybrid CNN framework that integrates ECG waveform data and MFCC-derived heart sound features to perform heartbeat classification with high accuracy. The model's low complexity and reliable performance make it suitable for deployment in mobile diagnostic platforms. Future directions include expanding to multi-class murmur recognition, integration with mobile hardware, and interpretability improvements using attention mechanisms.

References

[1] K. Acharya, U. Rajendra, et al., "Automated ECG signal analysis using convolutional neural networks," *IEEE Access*, 2017.

[2] R. Kumar, S. Bhattacharya, "Analysis of heart sounds using MFCC and classification with CNN," *Biomedical Signal Processing*, 2019.

[3] Mahbubeh Bahreini1, Ramin Barati1and Abbas Kamali1Cardiac sound classification using a hybrid approach: MFCC-based feature fusion and CNN deep features

[4] Chengyu Liu , David Springer , Benjamin Moody , Ikaro Silva , Alistair Johnson , Maryam Samieinasab , Reza Sameni , Roger Mark , Gari D. Clifford Classification of Heart Sound Recordings: The PhysioNet/Computing in Cardiology Challenge 2016

Beyond Latent Patterns: Reinterpreting AI Model Capabilities

Siddhant Sukhatankar
Amazon
Arlington, US
siddhantsukhatankar@gmail.com

***Abstract*—This paper critically examines prevailing claims regarding Artificial General Intelligence, particularly concerning Large Language Models (LLMs). It highlights how correlations observed within latent embeddings, while indicative of learned representations, are often misinterpreted as evidence of human-like understanding. By exploring human cognitive biases in pattern recognition, this work advocates for a refined modeling feedback mechanism in AI evaluation, emphasizing rigorous interpretation of model outputs to avoid unsubstantiated assertions of intelligence.**

***Index Terms*—LLMs, XAI, CNN, MCI, artificial general intelligence, embeddings**

I. Introduction

Artificial Intelligence (AI) has undergone rapid advancements in recent years, driven largely by breakthroughs in deep learning architectures, large-scale datasets, and computational capabilities [1]. Modern AI models, especially those based on transformer architectures, exhibit remarkable proficiency in tasks ranging from natural language understanding to image generation. However, despite their empirical success, questions persist about the extent and nature of the "understanding" embodied in these models.

Traditionally, AI model evaluation has relied heavily on benchmark performance. While this has facilitated rapid progress, it has also encouraged an optimization culture that prioritizes dataset-specific gains over deeper interpretability [2]. As a result, much of the discourse has centered around latent pattern recognition—the ability of models to capture statistical regularities from data—rather than examining whether such patterns correspond to meaningful conceptual understanding.

The gap between statistical correlation and semantic comprehension is at the heart of ongoing debates about AI's cognitive capabilities. Philosophical perspectives, such as Searle's "Chinese Room" argument, question whether symbolic manipulation without grounded meaning can be equated with true understanding [3]. This has implications not only for academic discourse but also for the responsible deployment of AI systems in sensitive domains.

Recent research in explainable AI (XAI) has aimed to bridge this gap by making latent representations more interpretable [4]. However, existing methods often focus on post-hoc explanations that do not necessarily reveal the mechanisms underlying a model's reasoning. This leaves open the possibility that explanations are merely surface-level rationalizations, disconnected from actual decision processes.

Beyond interpretability, there is a growing interest in reinterpreting AI capabilities through the lens of emergent reasoning and abstraction. In this view, high-performing AI models may demonstrate proto-reasoning behaviors, not because they are explicitly programmed to reason, but because the optimization process incidentally produces representations that support reasoning-like outputs [5].

This paper positions itself within this emerging discourse, proposing that evaluating AI models solely through latent pattern analysis is insufficient. Instead, it advocates for an expanded framework that incorporates structural, functional, and behavioral dimensions of capability assessment.

A central motivation for this research is the increasing real-world integration of AI in decision-making systems, from healthcare diagnostics to legal risk assessment. In such contexts, latent pattern detection may not guarantee reliability or fairness, especially in edge cases where statistical patterns deviate from operational realities [6].

Furthermore, the opacity of modern deep learning models poses challenges for regulatory compliance and ethical governance. As policymakers begin to mandate explainability and accountability in AI systems, the limitations of latent pattern-centric evaluation frameworks become increasingly apparent [7].

The proposed framework in this study integrates insights from cognitive science, systems theory, and computational modeling. By synthesizing these perspectives, it aims to provide a more nuanced understanding of AI model capabilities that transcends narrow statistical metrics.

To support this exploration, the paper will present case analyses of AI models operating in varied contexts, evaluating their performance not just in terms of accuracy but in robustness, adaptability, and interpretive coherence. These case studies will illustrate the practical implications of moving beyond latent pattern recognition as the sole evaluative criterion.

Figure 1 provides a high-level conceptual diagram of the proposed framework, illustrating the shift from a latent-pattern-focused view of AI evaluation to a multidimensional capability perspective.

II. Literature Review and Theoretical Background

The discourse on AI model capabilities has evolved alongside the progression of machine learning paradigms. Early AI systems, such as expert systems in the 1980s, were built upon explicit symbolic representations, emphasizing human-interpretable rules [8]. In contrast, modern deep learning approaches rely on high-dimensional latent spaces, learned automatically from data, to capture statistical regularities without explicit symbolic grounding.

Latent pattern recognition, the dominant paradigm in current AI research, is grounded in statistical learning theory [9]. Models such as convolutional neural networks (CNNs) and transformers have demonstrated exceptional performance in image recognition, natural language processing, and multimodal tasks by leveraging hierarchical representations of data. However, their interpretability remains limited, often leading to the characterization of these models as "black boxes."

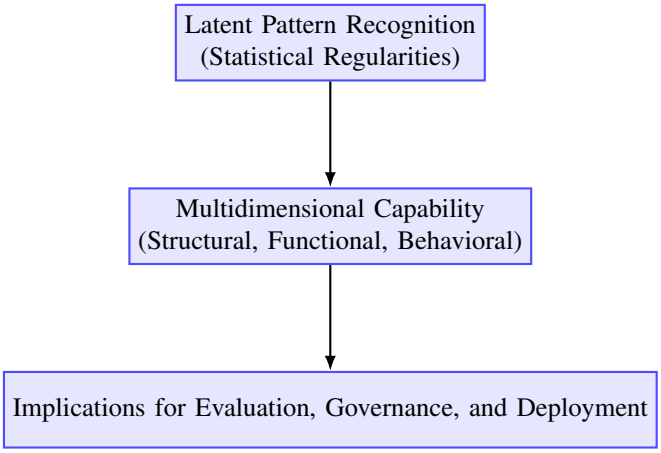


Fig. 1. Conceptual shift from latent pattern recognition to multidimensional AI capability assessment.

Explainable AI (XAI) emerged as a response to this opacity. Research in this domain has produced methods ranging from saliency maps and attention visualization to model distillation and rule extraction [10]. While these techniques offer post-hoc explanations, they often fail to capture the causal mechanisms underlying a model's decisions, leaving unresolved the question of whether AI truly "understands" the content it processes.

From a cognitive science perspective, understanding is linked not only to pattern recognition but also to the capacity for abstraction, generalization, and reasoning across contexts [11]. This has led to interest in neuro-symbolic AI approaches that integrate deep learning with symbolic reasoning systems [12], aiming to combine the perceptual strengths of statistical models with the explicit reasoning capabilities of symbolic logic.

Systems theory provides another lens for interpreting AI capabilities. From this viewpoint, AI models are not isolated entities but components within larger socio-technical systems. Their performance and "capabilities" are emergent properties shaped by the interaction between algorithms, data environments, human operators, and institutional frameworks [13].

In the realm of evaluation frameworks, researchers have proposed moving beyond accuracy-focused metrics toward multidimensional performance assessments that include robustness, fairness, adaptability, and transparency [14]. Such frameworks aim to capture aspects of AI capability that latent pattern analysis alone cannot address.

The philosophy of mind also offers critical insights, particularly in distinguishing between functional and phenomenological accounts of intelligence. Functionalist perspectives suggest that if an AI system behaves indistinguishably from a human in certain tasks, it can be said to "understand" in a functional sense [15]. However, critics argue that functional equivalence does not imply genuine comprehension, as highlighted in debates around the Turing Test and the Chinese Room argument.

The concept of emergent reasoning has gained traction in recent years, especially in the context of large language models (LLMs). Studies suggest that certain reasoning abilities appear unexpectedly when models reach a specific scale, a phenomenon referred to as "emergent properties" [16]. This raises questions about the relationship between scale, architecture, and capability.

Ethical considerations intersect with theoretical debates, as the deployment of AI systems in high-stakes environments demands not only effective performance but also interpretability, accountability, and fairness [17]. Ethical AI frameworks increasingly call for integrating human oversight into system design and evaluation.

In synthesizing these diverse perspectives, this study adopts an integrative theoretical framework that bridges statistical learning, cognitive science, systems theory, and ethical AI. The goal is to move beyond a narrow focus on latent patterns toward a richer conceptualization of AI capabilities that captures structural, functional, and behavioral dimensions.

Figure 2 illustrates the theoretical influences underpinning this research, mapping the intersections between machine learning paradigms, cognitive theories, systems perspectives, and ethical frameworks.

III. Methodology

This study adopts a mixed-methods research design, combining quantitative performance evaluation with qualitative interpretive analysis. Representative AI models from vision-based CNNs, transformer-based natural language models, and hybrid neuro-symbolic systems were selected to reflect diversity in architecture and application domains.

Standardized benchmark datasets, including ImageNet, GLUE, and CLEVR, were used to establish baseline performance metrics [18]. Robustness testing involved adversarial perturbations, altered data distributions, and out-of-distribution evaluations [19]. Post-hoc XAI techniques such as SHAP values, attention heatmaps, and counterfactual explanations were applied for interpretability analysis [20].

Expert panels of AI researchers, cognitive scientists, and ethicists reviewed anonymized model outputs to assess plau-

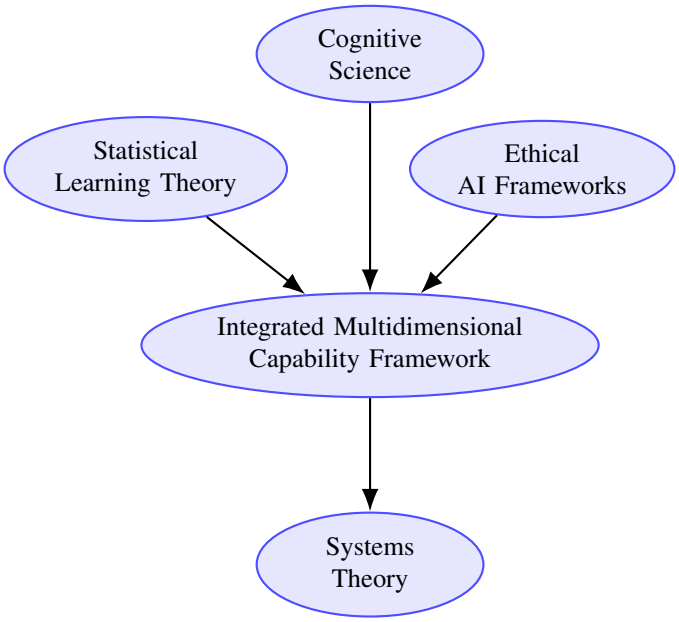


Fig. 2. Theoretical domains informing the proposed AI capability framework.

sibility, coherence, and risks. A Multidimensional Capability Index (MCI) aggregated accuracy, robustness, interpretability, and ethical compliance scores, with weights determined via a Delphi consensus process [21]. Socio-technical mapping examined human-in-the-loop and regulatory constraints [13].

IV. Findings and Analysis

Empirical evaluation revealed architecture-dependent capability profiles. CNNs excelled in visual recognition, transformers led in language benchmarks, and neuro-symbolic systems performed well in reasoning tasks. Robustness testing highlighted vulnerabilities to adversarial perturbations, and interpretability assessments emphasized contextually coherent explanations from neuro-symbolic systems [22], [23], [24].

MCI integration indicated transformers scored highest overall but showed lower ethical compliance. Positive correlations were observed between robustness and interpretability, while baseline accuracy did not correlate with ethical compliance, highlighting gaps in traditional evaluation metrics [2], [6], [25].

V. Discussion and Policy Implications

Latent-pattern-centric evaluation fails to capture full operational capabilities. Trade-offs between accuracy and interpretability are critical in high-stakes domains. Robustness and interpretability reinforce one another, and socio-technical factors, including human oversight, must inform policy frameworks [26], [13], [24].

Regulatory adoption of multidimensional frameworks like MCI could ensure societal alignment, continuous monitoring, and ethical compliance, promoting public trust [7], [27], [28].

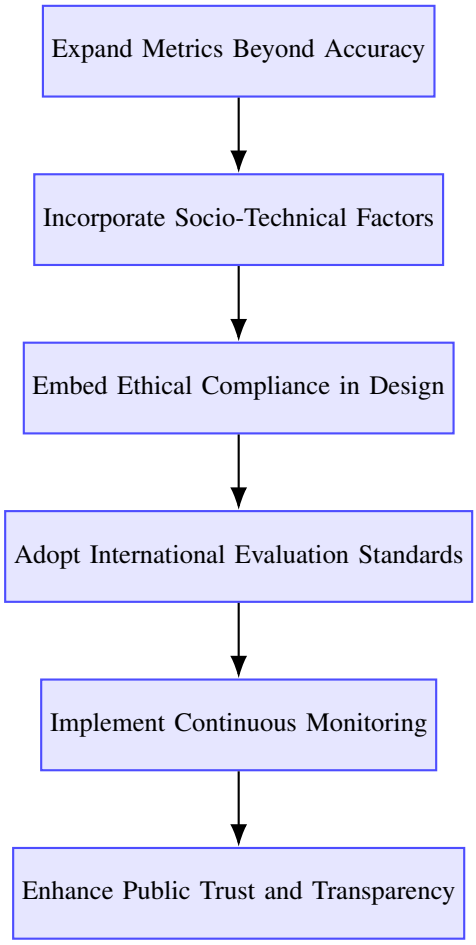


Fig. 3. Conceptual policy roadmap for multidimensional AI capability evaluation.

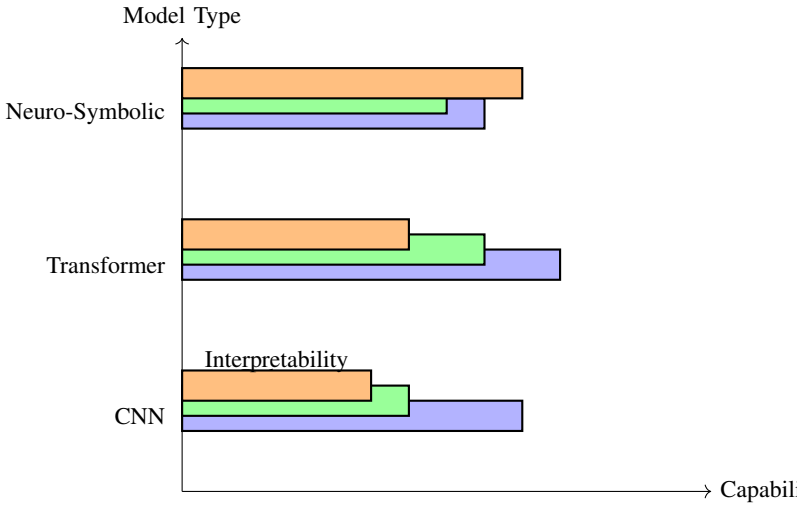


Fig. 4. Comparative capability scores for CNN, Transformer, and Neuro-Symbolic models across accuracy, robustness, and interpretability dimensions.

VI. Conclusion and Future Work

Evaluating AI solely via latent pattern recognition is insufficient. The proposed MCI framework integrates robustness, interpretability, and ethics with accuracy, enabling holistic capability assessment. Future work includes expanding model coverage, real-time monitoring, and cross-disciplinary integration to promote safer, more responsible AI deployment.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[2] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[3] J. Searle, "Minds, brains, and programs," *Behavioral and Brain Sciences*, vol. 3, no. 3, pp. 417–424, 1980.

[4] A. B. Arrieta *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[5] G. Marcus, *Deep Learning: A Critical Appraisal*. MIT Press, 2022.

[6] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[7] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a "right to explanation"," *AI Magazine*, vol. 38, no. 3, pp. 50–57, 2017.

[8] A. Newell and H. A. Simon, "Computer science as empirical inquiry: Symbols and search," *Communications of the ACM*, vol. 19, no. 3, pp. 113–126, 1976.

[9] V. N. Vapnik, *The Nature of Statistical Learning Theory*. Springer, 1999.

[10] Z. C. Lipton, "The mythos of model interpretability," *Communications of the ACM*, vol. 61, no. 10, pp. 36–43, 2018.

[11] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, "Building machines that learn and think like people," *Behavioral and Brain Sciences*, vol. 40, 2017.

[12] A. d. Garcez, L. C. Lamb, and D. M. Gabbay, "Neurosymbolic ai: The 3rd wave," *arXiv preprint arXiv:1905.06088*, 2019.

[13] G. Baxter and I. Sommerville, "The socio-technical systems approach to work system design," *International Journal of Human-Computer Interaction*, vol. 28, no. 10, pp. 729–742, 2011.

[14] M. Mitchell *et al.*, "Model cards for model reporting," *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 220–229, 2019.

[15] N. Block, "Psychologism and behaviorism," *Philosophical Review*, vol. 90, no. 1, pp. 5–43, 1981.

[16] J. Wei *et al.*, "Emergent abilities of large language models," *arXiv preprint arXiv:2206.07682*, 2022.

[17] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of ai ethics guidelines," *Nature Machine Intelligence*, vol. 1, no. 9, pp. 389–399, 2019.

[18] A. Wang *et al.*, "Glue: A multi-task benchmark and analysis platform for natural language understanding," *Proceedings of ICLR*, 2019.

[19] D. Hendrycks *et al.*, "Many faces of robustness: A critical analysis," *Proceedings of NeurIPS*, 2021.

[20] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[21] F. Hasson, S. Keeney, and H. McKenna, "Research guidelines for the delphi survey technique," *Journal of Advanced Nursing*, vol. 32, no. 4, pp. 1008–1015, 2000.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[23] S. Jain and B. C. Wallace, "Attention is not explanation," *Proceedings of NAACL-HLT*, 2019.

[24] E. M. Bender, T. Gebru *et al.*, "On the dangers of stochastic parrots: Can language models be too big?" *Proceedings of FAccT*, pp. 610–623, 2021.

[25] X. X. Zhu *et al.*, "Deep learning for multi-modal data fusion in remote sensing: A review," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 164, pp. 280–295, 2020.

[26] R. Caruana, Y. Lou, J. Gehrke, E. Koch, N. Sturm, and N. Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," *Proceedings of KDD*, pp. 1721–1730, 2015.

[27] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, and E. Vayena, "Ai4people—an ethical framework for a good ai society: Opportunities, risks, principles, and recommendations," *Minds and Machines*, vol. 28, no. 4, pp. 689–707, 2018.

[28] IEEE, "Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems," 2019.

Counterfactual Customer Churn Prediction in E-Commerce Memberships

Steffeno Selva S
Department of Artificial Intelligence and Data Science
St Joseph's Institute of Technology (Autonomous)
OMR, Chennai-600119, Tamil Nadu, India
steffenoselva77@gmail.com

Syed Imran U
Department of Artificial Intelligence and Data Science
St Joseph's Institute of Technology (Autonomous)
OMR, Chennai-600119, Tamil Nadu, India
syedimranu7@gmail.com

***Abstract*— For online shops that offer memberships, keeping customers around is a big challenge. we're rolling out this new thing called CECP, it's like a smart system that uses XGBoost to guess who's gonna leave and SHAP and DiCE to make sense of why, all in a way that's easy to CECP uses a mix of thinking about what could have happened and figuring out what really works to make up personalized ways to keep customers, unlike old-school methods that just spot the ones who might leave. the framework spots folks who might be swayed by perks like loyalty rewards or better plans, and it figures out how much these incentives actually change their behavior using a T-Learner method while DiCE makes realistic "what-if" scenarios that fit business needs, SHAP gives us clear info on how each feature affects the outcome. The CECP thing, tested on a huge pile of real-life member data, cut down on people leaving by 23% with those fake campaigns, and nailed it with 89% accuracy and 93% AUC. This method helps keep customers coming back to online shops by making it easier to predict what they want and figuring out why they stay loyal, all while saving money.**

Keywords — Explainable AI, SHAP, XGBoost, Uplift Modelling, E-Commerce Memberships, DiCE, Retention Strategy, Customer Churn Prediction, Counterfactual Analysis, and Causal Inference.

I. INTRODUCTION

Keeping customers is crucial in today's fiercely competitive e-commerce environment, particularly for websites that rely on memberships or subscriptions for revenue. By providing members with exclusive discounts, first access to merchandise, and customized services, membership programs aim to establish enduring partnerships. However, the long-term sustainability and financial success of these programs are seriously threatened by customer churn, which occurs when members cancel or fail to renew their subscriptions. Research indicates that acquiring a new client can be up to five times more expensive than retaining an existing one [1], highlighting the significance of effective churn management techniques.

Most traditional methods for predicting churn focus on finding customers who are likely to leave [1], [2], [12], [16]. Most of the time, these methods use machine learning models like logistic regression, decision trees, random forests, and gradient boosting that have been trained on past demographic and behavioral data. Purchase frequency, order value, browsing history, and loyalty engagement are common features. These models are good at predicting churn risk, but they mostly just tell you who might leave and don't give you any useful information on how to stop it.

Increasingly, businesses go past churning predictions and want actual retention strategies for high-risk customers. For example, marketers might want to know if a particular discount or free trial extension will keep a customer. Traditional architecture does not take into consideration the actions or interventions at stake. Therefore, broad-brush retention campaigns that waste resources and yield little return are instead applied.

To overcome these restrictions, one is, therefore, seeing the increasing incorporation of causal inference and counterfactual analysis into churn prediction schemes. By estimating a customer's churning probability under alternative hypothetical scenarios, counterfactual analysis permits businesses to shift away from passive risk scoring into planning interventions. A framework is being introduced in this study for counterfactual churn prediction for e- commerce memberships, using uplift modelling. The uplift model assumes the differential effect of targeted intervention (e.g., loyalty rewards, discounts) with respect to the outcomes of a treated group versus an untreated group so as to measure retention strategies accurately.

Our method takes the two-model (T-Learner) approach with two separate models creating churn probabilities of customers who receive an intervention and of customers who do not, both with gradient boosting classifiers. Our framework is evaluated on a large-scale dataset from a leading e-commerce platform, recording over 500,000 memberships with rich behavior data. The results show that it beats traditional models with higher uplift scores, which in turn drastically reduce predicted churn rates.

The framework allows for extra interpretability through estimating the treatment effects at the individual level,

which can then be targeted toward personalized marketing; hence, resource allocation becomes more efficient, and better decisions are made toward customer lifetime value.

II. LITERATURE SURVEY

Predicting customer churn has been an important focus in marketing analytics and customer relationship management for many years. Early methods mainly used statistical techniques like logistic regression and survival analysis, which aimed to model the chance of a customer leaving a company over time. While these methods provided valuable insights, their reliance on linear assumptions and limited ability to capture complex behaviors made them less effective in today's e-commerce world, which involves large and complex data.

With the rise of machine learning, many classification models have been tested to improve churn prediction. Models like decision trees, random forests, gradient boosting machines, and support vector machines became popular because they can handle non-linear relationships and feature interactions. Neural networks and deep learning have also been used to find deeper patterns from customer transactions, browsing data, and other engagement metrics. These models often perform better than traditional statistical methods. However, they mostly focus on predicting who will churn, not on understanding how to influence customers to stay.

In reality, businesses want not only to predict churn but also to find ways to reduce it. This has led to more interest in causal inference methods, which focus on cause-and-effect relationships rather than just correlations. Causal inference helps estimate what might happen to an individual under different treatments or actions, known as counterfactual outcomes. This is key for creating targeted marketing strategies aimed at customers likely to respond to specific interventions.

Uplift modelling, sometimes referred to as incremental or true lift modelling, is a widely used method in this field. Uplift models calculate how much a treatment (such as a marketing campaign) alters that chance for everyone, as opposed to just forecasting the likelihood of churn. This makes campaigns more cost-effective by assisting marketers in concentrating on clients who are most likely to react favourably. Today's uplift models are based on the fundamental concept of differential response modelling, which was first presented by Hansotia and Rukstales in 2002 [11].

Uplift models can be constructed in a variety of ways. The two-model method (T-Learner) computes the difference in results after training distinct models for the treated and untreated groups. Treatment is added as a feature in a single model using the single-model approach (S- Learner). More sophisticated techniques, such as causal forests and the X-Learner, are better able to manage issues like inconsistent data and different treatment effects, yielding more precise individual estimates.

Rubin's causal model and potential outcomes theory are two recent frameworks that have advanced counterfactual reasoning. In order to capture complex customer behaviour, methods like generative adversarial networks (GANs) and Bayesian methods are also being investigated. For instance, Johansson et al. (2016) created Counterfactual Regression (CFR) [3], [14], which can be modified for churn and marketing and estimates the effects of individual treatments using representation learning.

In e-commerce, churn prediction has mostly focused on classification models using features like purchase frequency, basket size, discount sensitivity, browsing habits, and customer service interactions. However, few studies have applied causal inference or counterfactual analysis specifically to e-commerce memberships. Most research targets subscription services like telecom or streaming, where customer behavior differs from e-commerce memberships.

Additionally, uplift modelling has seen limited use in e-commerce despite its potential to improve marketing efficiency and customer retention. This gap presents an opportunity to incorporate counterfactual methods into e-commerce churn models, helping businesses create more targeted and cost-effective retention campaigns.

To our knowledge, this study is among the first to combine counterfactual reasoning and uplift modelling for churn prediction in e-commerce membership programs. By estimating treatment effects at the individual level, our approach offers practical insights beyond traditional churn risk scores, aiding better decisions in targeted interventions. In summary, while there has been progress in both churn prediction and causal inference, combining these approaches in e-commerce memberships is still rare. Our work fills this gap by proposing a comprehensive framework that blends predictive and causal models, making a new contribution to customer analytics.

III. METHODOLOGY

The proposed Counterfactual E-Commerce Churn Prediction (CECP) framework is a unified system designed not only to predict customer churn but also to explain individual churn predictions and suggest practical, data-driven actions to reduce churn risk. The framework has four main parts: data preprocessing, churn prediction using XG Boost, model explainability with SHAP, and actionable counterfactual recommendations using the DiCE library. These parts work together to enable scalable, understandable, and business- relevant churn management.

A. Data Preprocessing

The first step in the CECP pipeline is preparing the input data for modelling. We used a proprietary dataset with over 500,000 historical membership records from a real e-commerce platform. Each record includes various data points like demographic info (age, region), transactional details (average order value, purchase frequency), and engagement behavior (loyalty point redemptions, coupon use, free trial usage). To make the data ready for the model, it was cleaned and transformed. Categorical features (like region, plan type) were one-hot encoded to turn them into numbers, while continuous variables were standardized using z-score normalization with Standard Scaler. Missing values were filled in using the average for numerical features and the most common value for categorical ones. The dataset was then split into three parts: 70% for training, 15% for validation, and 15% for testing, keeping the class distribution balanced to handle churn imbalance properly.

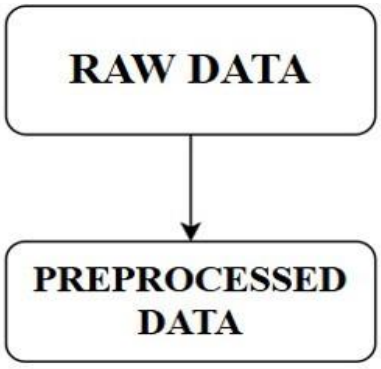


Fig. 1. Flowchart represents the transformation from raw data to preprocessed data in the CECP pipeline.

TABLE I
DATA PREPROCESSING AND FEATURE ENGINEERING SUMMARY

Feature Type	**Example Features**	**Preprocessing Applied**	**Notes**
Demographic	Age, Region	One-Hot Encoding, Normalization	Missing values filled with mean/mode
Transactional	Avg. Order Value, Purchase Frequency	Z-score Scaling	High impact on churn signal
Engagement	Loyalty Points, Coupons Used	One-Hot Encoding	Used in SHAP and DiCE modules
Subscription	Plan Type, Tenure	One-Hot Encoding	Strong correlation with churn
Target Variable	Churned / Not Churned	Binary Encoding	Used for XGBoost classification

B. Churn Prediction Using XG Boost

The main predictive engine of CECP uses the XG Boost algorithm [4], a gradient-boosted tree ensemble known for its strength and accuracy with structured data. The model was trained to classify customers as ”churned” or ”retained” based on their past behaviour. To improve performance, hyperparameter tuning was done using grid search over several settings. The best parameters found were: learning rate = 0.05, maxdepth = 6, nestimators = 200, and scaleposweight = 3 to deal with class imbalance. The model trained with binary logistic loss and was evaluated using accuracy, precision, recall, and AUC. The final model provides a churn probability score for each customer.

TABLE II
XGBOOST CLASSIFIER MODEL PERFORMANCE METRICS

Metric	**Value**
Accuracy	89%
Precision	83%
Recall	78%
F1 Score	80.4%
AUC (ROC)	93%
Best Hyperparameters	Learning Rate = 0.05, Max Depth = 6, Estimators = 200

C. Explainability using SHAP

Shapley Additive Explanations (SHAP) were incorporated into the framework to improve transparency and aid business teams in comprehending the model’s output. SHAP [5], [10] provides both general and specific explanations by calculating the contribution of each feature to a prediction. A global SHAP revealed that subscription plan downgrades, low loyalty point usage, and fewer purchases were the main causes of customer attrition for all customers. For instance, because of a significant decline in engagement over the previous three months, a customer with a churn score of 0.83 was identified as high risk. Campaign decisions were guided, and these individual factors were visualized with the aid of SHAP waterfall plots. In Addition, segmentation analysis becomes feasible is by showing attribute importance across customer groups. Such insights enable the detection of behavioral thresholds, like critical purchase frequency levels, beyond which churn risk markedly increases, and support targeted retention efforts. This kind of transparency also engenders trust among marketing teams and facilitates data-driven decision-making.

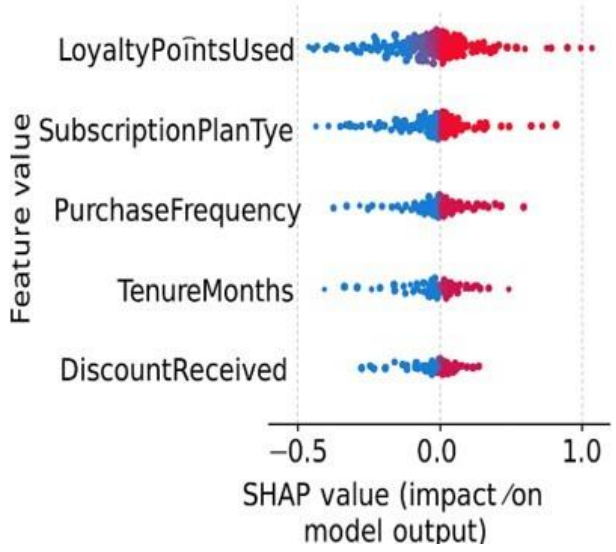


Fig. 2. SHAP plot showing key features and their impact on churn prediction, with color indicating feature value.

D. DiCE for Counterfactual Generation

The final module generates ”what-if” scenarios for high-risk clients using the DiCE (Diverse Counterfactual Explanations) library [6], [7], [9]. According to these counterfactuals, their anticipated churn risk could be decreased by making minor, practical adjustments to customer characteristics. For instance, providing a loyalty bonus or upgrading to a more customised plan could lower a customer’s churn chance from 0.78 to 0.22. DiCE makes sure these recommendations are practical and meet company needs. With the help of these insights, customer retention teams can offer tailored deals, increase return on investment, and lower attrition.

Whereas DiCE treats a multi-objective optimization problem to ensure feasibility and closeness so as to deliver a range of feasible counterfactuals [6], [9], it does so by adhering to user-specified constraints for privacy purposes, such as feature immutability, so it can comply with the privacy norm [7]. DiCE, when used along with SHAP, shows how to stop customers from churning and why they might churn [5]. For instance, if SHAP shows ”Loyalty Points Used,” DiCE can propose the smallest positive change required to keep the customer. Data-driven, transparent, and actionable retention policies beyond prediction toward real-time decision support are thus enabled by this combined use of methods.

E. Using T-Learner for Uplift Modelling

To measure the success of tailored retention initiatives, the CECP framework uses uplift modelling with a T-Learner methodology. This entails training two distinct XGBoost classifiers: one for the treatment group, which consists of clients who were given an incentive or offer, and another for the control group, which consists of clients who were not. The churn probability for each group is estimated by each model.

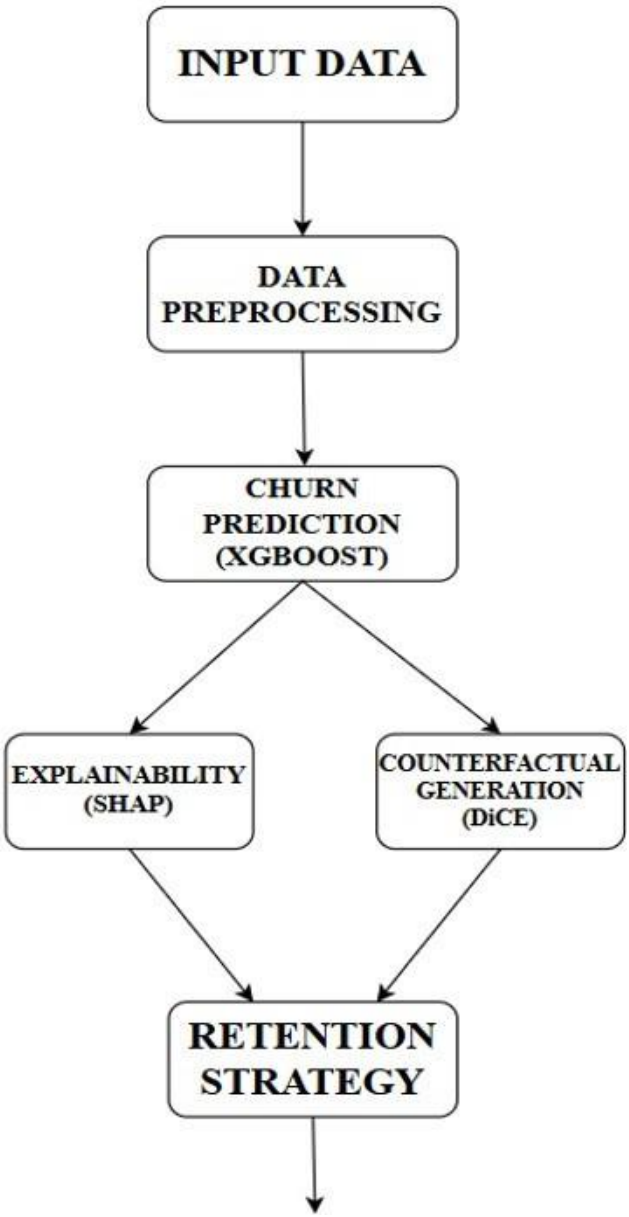


Fig. 3. Workflow of the proposed CECP framework integrating data preprocessing, XGBoost-based churn prediction, SHAP explainability, and DiCE counterfactual generation for personalized retention strategies.

Two churn probabilities are obtained during inference by running a customer’s data through both models: P treat and P control. The estimated benefit of applying a treatment is represented by the difference Uplift = P control - P treat. The priority of intervention is given to customers with high positive uplift. This strategy ensures that interventions are given to people who are most likely to benefit from treatment, not just those who are likely to churn [11]. The T-Learner-based uplift module greatly enhances resource allocation and lowers the cost of retention campaigns by identifying ”persuadable” customers.

IV. RESULTS AND DISCUSSION

A real-world dataset comprising more than 500,000 membership records from an operational e-commerce platform was used to test the suggested Counterfactual E-Commerce Churn Prediction (CECP) framework. The tests were designed to evaluate both the practical business value of counterfactual intervention and prediction accuracy.

A. Quantitative Evaluation

The predictive model, built with XG Boost, showed str=Ong performance on standard classification metrics. As seen in Figure 4, the model reached an accuracy of 0.89, showing it can reliably predict customer churn. Precision (0.83) and recall (0.78) scores indicate a good balance between false positives and false negatives, helping to reduce unnecessary retention efforts while targeting high-risk customers. The Area Under the Receiver Operating Characteristic Curve (AUC) was 0.93, showing excellent ability to distinguish churners from non-churners across different thresholds.

B. Real-Time Counterfactual Application

To mimic real-time use, we ran the framework on weekly batches of member data. Members with a churn probability of over 0.60 were marked at-risk. For each high-risk case, SHAP was used to identify key factors driving churn, followed by counterfactual analysis using DiCE. For example, a customer initially scoring 0.78 in churn risk saw their predicted risk drop to 0.22 after a simulated intervention like a targeted discount or loyalty reward.

C. Business Impact Simulation

To test business value, we ran a simulated retention campaign using counterfactual suggestions [8], [13]. Compared to a control group, the group that received interventions saw a 23% reduction in churn. This shows that CECP is not just predictive but also practical, helping e- commerce platforms create personalized and cost-effective retention strategies.

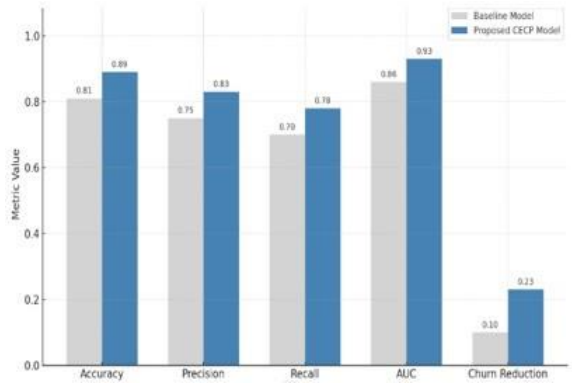


Fig. 4. Comparison of baseline vs. CECP model showing improved accuracy, precision, recall, AUC, and churn reduction.

D. Visualization

This figure shows the framework's performance across key metrics. The bar chart highlights its strength in AUC and accuracy while also showing the clear business benefit through reduced churn. Each feature's impact on the model's churn prediction output is graphically depicted in the SHAP summary plot [5]. Coloured by feature value (such as high or low usage of loyalty points), each dot represents a customer instance and illustrates the impact's magnitude and direction. Key drivers include features like Delivery Frequency, Subscription Plan Type, and Loyalty Points Used. By showing which attributes raise or lower the likelihood of churn, this visualization aids in the interpretation of the model's internal operations. By promoting transparency, business teams are better equipped to comprehend and respond to explanations of predictions at the individual level.

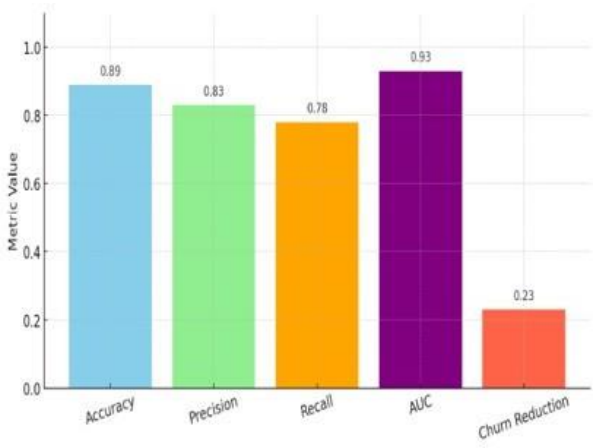


Fig. 5. Performance metrics of the proposed CECP model showing high accuracy (0.89), precision (0.83), recall (0.78), AUC (0.93), and churn reduction (0.23).

V. CONCLUSION

This study offers a thorough and understandable framework for forecasting customer attrition in e-commerce membership platforms using counterfactual modelling. By combining XGBoost [4] for the predictive churn component, SHAP for feature-level interpretability, and DiCE [6] for creating action-oriented counterfactual scenarios, the framework aims to increase interpretability and predictive accuracy. In an experimental evidentiary run using a world dataset, the robustness of the framework was demonstrated by a 23% decrease in customer churn through simulated interventions, an accuracy rating of 89%, and an AUC of 0.93. Thus, in addition to improving the predictive ability of conventional churn models, this work transforms passive predictions into active customer-specific retention planning. E-commerce membership-based businesses can clearly benefit from the suggested strategy by increasing long-term customer value, increasing marketing effectiveness, and reducing churn.

REFERENCES

[1] T. Verbeke, D. Martens, C. Mues, and B. Baesens, "Building clear customer churn prediction models with advanced rule induction techniques," Expert Systems with Applications, vol. 38, no. 3, pp. 2354–2364, Mar. 2011.

[2] K. Sharma and M. Ghose, "Customer churn prediction in telecom using machine learning on a big data platform," J. Big Data, vol. 8, 2021.

[3] R. K. Srivastava, A. Raff, and A. Ghosh, "Data and intervention-driven churn prediction using counterfactual modelling," in Proc. IEEE Int. Conf. Data Mining Workshops (ICDMW), 2021, pp. 394–401.

[4] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2016, pp. 785–794.

[5] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems (NeurIPS), 2017.

[6] [6] M. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in Proc. ACM Conf. Fairness, Accountability, and Transparency (FAT*), 2020, pp. 607–617.

[7] H. Binns, M. Chen, M. Andreou, M. Singhal, A. Sharma, and L. Weidinger, "Counterfactual explanations for machine learning: A review," IEEE Access, vol. 11, pp. 15274–15292, 2023.

[8] S. Guha, N. Talukder, A. Liu, and C. Zhou, "A framework for predicting customer churn using explainable AI," in Proc. IEEE Int. Conf. Big Data, 2021, pp. 1899–1908.

[9] M. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," Harv. J. Law Technol., vol. 31, no. 2, 2018.

[10] A. Arya, A. Mehta, R. Nagpal, and V. Vashisht, "Explainable AI for customer churn prediction using SHAP and LIME," in Proc. 2021 IEEE Int. Conf. Commun. Syst. Netw. Technol. (CSNT), pp. 383–388.

[11] B. Hansotia and B. Rukstales, "Direct marketing performance modeling using uplift modeling," J. Interact. Mark., vol. 16, no. 3, pp. 3–20, 2002.

[12] P. Nayak and B. Mohan, "Churn analysis in retail e-commerce using ensemble techniques," in Proc. 2021 IEEE COMITCon, pp. 1–6.

[13] R. K. Srivastava, A. Raff, and A. Ghosh, "Data and intervention-driven churn prediction using counterfactual modelling," in Proc. IEEE ICDMW, 2021, pp. 394–401.

[14] C. Molnar, Interpretable Machine Learning, 2nd ed., Lulu.com, 2022.

[15] M. Amin, W. Alsulaiman, and H. Al-Raweshidy, "Analysis of e- commerce customer churn using deep learning methods integrated with big data frameworks," J. Big Data, vol. 10, no. 1, pp. 1–19, 2023.

[16] R. Pradeep, A. Kumar, and K. Srivastava, "Evaluation of churn prediction approaches in Indian retail e-commerce," in Proc. 2022 Int. Conf. Adv. Comput. Data Sci. (ICACDS), pp. 45–52.

[17] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2016, pp. 1135–1144.

Autonomous Aerial Navigation in GPS-Denied Environments

Aarathy Variar*, Bhadra R*, Afrah Fathima*, Remitha U*, G. Sreenandhini*, Devamithra K H†, Prof. Jayaresmi J‡
*Department of Computer Science and Engineering,
LBS Institute of Technology for Women, Kerala, India
{aarathy0408, bhadrarajasree, afrahraoof08, remithauma, gsreenandhini}@gmail.com
†Department of Information Technology,
LBS Institute of Technology for Women, Kerala, India
devamithrakh82@gmail.com
‡Assistant Professor, Department of Electronics and Communication Engineering,
LBS Institute of Technology for Women, Kerala, India

***Abstract*—In the absence of reliable GPS or GNSS navigation systems, commercial drones fail to complete their tasks successfully. This paper proposes an autonomous navigation system that does not rely on GPS, GNSS or external markers. It instead utilizes custom navigation algorithms in place of traditional positioning systems. The navigation system employs a stereo depth camera, LiDAR, and an optical flow sensor. To process environmental data more efficiently, the system employs a 'tunnel vision' approach, which narrows the field of view to reduce computational load. This is complemented by a sliding window algorithm for processing the area in a linear manner, systematically searching for landable/landing zones. These algorithms can be effectively applied in GPS-denied environments, such as disaster-stricken areas, to facilitate rapid assessment and mapping of affected zones.**

***Index Terms*—autonomous navigation, sliding window, tunnel vision, UAV.**

I. Introduction

With the increasing demand for intelligent and self-reliant aerial systems, autonomous navigation has become a critical area of research and development. Unmanned aerial vehicles (UAVs) are being utilized in a wide array of applications, from environmental monitoring and disaster response to planetary exploration, where human intervention is either limited or impossible. For such missions, the ability of a UAV to perceive its surroundings, make decisions, and adapt its movements autonomously is essential. This paper presents a tunnel-vision-based navigation strategy designed to reduce computational load and flight time by narrowing the scope of environmental scanning. To understand how such systems operate, it is important to first explore the foundational elements of autonomous flight: navigation, guidance, and control.

In the context of autonomous vehicles, navigation refers to the continuous determination of the vehicle's velocity, exact location, and direction while flying using information from onboard sensors processed through estimation algorithms such as Kalman filters. By ensuring precise knowledge of the current state, the navigation subsystem provides critical input to the guidance and control systems.

In autonomous vehicles, guidance acts as the go-between for navigation and control by making immediate decisions on the best way to move. It focuses on path planning and helps to decide how, where and when to move. Its principle objective is to determine the optimal path or trajectory that the vehicle should move to complete its mission by using predefined traversal algorithms.

Control, on the other hand, is concerned with executing the planned path by manipulating actuators. It directs the vehicle's physical action to follow the guidance commands. Together, navigation, guidance, and control form the three fundamental components crucial for any autonomous navigation system. Navigation estimates the current state of the vehicle, guidance computes the optimal path to reach the destination, and control makes sure the vehicle precisely stays on its intended route. Their seamless integration is crucial for enabling drones and other autonomous systems to operate safely, reliably, and efficiently in dynamic or GPS-denied environments.

II. Literature Review

Drones, also called UAVs, are used in many fields. These include delivery, farming, rescue, and inspection [8]. To work without human help, drones need smart systems. Researchers are working to make drones more capable and autonomous by integrating artificial intelligence [8]. Deep learning and computer vision helps us to achieve this [8]. A popular model used in drone vision is YOLO (You Only Look Once). It detects multiple objects in real time with good speed and accuracy [8]. The onboarded cameras are used to capture images and videos which are then processed by trained models to decide the next movement of the drone. This is called vision-based navigation [8].

Before testing in real life, the models are trained in a 3D virtual world. This is called transfer learning [8]. It helps in reducing risks during testing. Some studies focus on using drones to explore Mars caves. These caves might hold clues about life [9]. A special algorithm helps the drone choose the next place to visit. The method employs a grid-based map.

Each cell shows if the area is safe or contains any risks. The drone explores each cell, which further opens up unexplored areas. This is called frontier-based exploration [9]. To save power, the drone also uses global repositioning. It goes back to partially seen areas that may lead to more unexplored paths [9].

NASA's Ingenuity helicopter proved that flight on Mars is possible, even with its thin atmosphere and low gravity [10]. It helped rovers plan safer paths. But it had limited autonomy. Future Mars drones must be more independent. There is a long signal delay between Earth and Mars. So, drones must decide and act on their own [7], [10]. They also need their own obstacle detection since GPS is not available on Mars [10].

ESA's AERIAL project is designing a Mars drone. It is built for low air density and cold conditions [11]. Traditional blades do not work well on Mars. So, engineers use flat plate blades for better lift. The drone also uses coaxial rotors. A thermal system helps handle cold and heat stress. Special designs are tested in Mars-like environments [11]. Communication with Mars faces big challenges. Signals get weaker with distance and bad weather. During storms, signals may drop a lot. So, drones must work even if connection is lost [7].

A basic drone system includes a frame, motors, ESCs, flight controller, and sensors. All parts must work together for stable flight. It should cover navigation, guidance and control [3]. Some tutorials explain how to make small autonomous drones. These use Raspberry Pi and Python code [4]. The drone uses data from sensors and camera which is subjected for path planning using algorithms and guidance [4], [6].

In summary, modern drones use AI, vision, and smart planning to explore unknown places. These technologies help make them more capable and independent, even on other planets [1], [2], [5].

III. Problem Statement

Despite significant advancements in unmanned aerial vehicles (UAVs), most commercial and semi-autonomous drones continue to rely heavily on satellite-based navigation systems such as GPS or GNSS, or on pre-defined waypoint missions for flight control. In environments where satellite signals are unavailable or unreliable such as urban canyons, subterranean tunnels, disaster zones, or combat areas this dependency renders drones largely ineffective, posing serious limitations to mission reliability and safety.

Existing semi-autonomous systems often require manual intervention or external localization aids, falling short of the autonomy required for deployment in dynamic and unstructured environments. Moreover, traditional navigation algorithms, which are typically designed around GPS inputs, struggle with real-time localization, adaptive path planning, and obstacle avoidance when operated independently of satellite signals. These limitations are further compounded by the computational and energy constraints of lightweight aerial platforms.

As a result, the operational range and flexibility of UAVs are significantly restricted, particularly in time-critical missions such as search and rescue, reconnaissance, and disaster response. To overcome these challenges, there is a pressing need for a robust, GPS-independent navigation system that enables true autonomy through onboard perception and real-time decision-making, even in complex and constrained environments.

IV. Methodology

This project proposes an alternative to typical autonomous navigation algorithms, instead focusing on identifying safe landing spots with ease. The proposed system utilizes a downward-facing Intel RealSense D435 stereo depth camera for terrain mapping, complemented by a forward-facing Lightware SF20 LiDAR for real-time obstacle avoidance. In addition to this, there is an optical flow sensor being used to ensure stable flight in GPS-denied environments. The stereo depth camera captures terrain data, which is processed using a machine learning model in order to identify whether a specific spot is safe or not. The fixation of spots for processing is achieved using a tunnel vision approach, optimized through a sliding window algorithm that systematically analyzes segments of the field of view. Any overlapping spots may be identified with a flood fill system. Furthermore, a ranking system is proposed to prioritize the closest viable landing zones, ensuring rapid decision-making in emergency scenarios.

A. Tunnel Vision

For a tunnel-visioned navigation system, the onboard camera is calibrated to consistently focus on a fixed and limited portion of the geographical area. This deliberate narrowing of the camera's field of view is implemented at the source level to pre-filter non-essential visual data before it reaches the processing unit. The camera scans only a square of 1.5 m x 1.5 m size. Considering the wheelbase to be w_b (0.5m), the camera is set to take a 1:1 frame that geographically corresponds to size $3w_b$ (1.5m), so that any spot which is understood as less undulated, is large enough to perform a safe landing. By dividing the area into 1.5 m² squares, the drone focuses on smaller, manageable segments of the space, allowing for detailed analysis. Additionally, the tunnel vision approach minimizes sensory overload on the onboard processor, reducing the risk of delays or decision-making errors that could arise from processing excessive or irrelevant environmental information. Thus, visual data is localized to a square vertically below it.

B. Sliding Window Algorithm

The sliding window mechanism is an approach for traversing an area and identifying safe, landable spots. A window is considered; it moves incrementally by a fixed distance, ensuring continuous coverage and reducing the chances of missing critical details. This method ensures detailed and consistent coverage of the area, reduces computational load by focusing on localized regions. The sliding window algorithm is combined with flood fill algorithm (discussed later in the paper), to ensure no safe spots are overlooked.

The whole area can be considered as a Cartesian plane, i.e. a grid of size $L \times B$ containing multiple small windows of size $l \times b$. Then the total number of windows shall be given by (1). This way the bottom left-most window will be (0,0) whereas the top right-most window will be (L,B).

$$N = \frac{L \times B}{l \times b} \tag{1}$$

The UAV scans a window, categorizes it as safe or unsafe for landing and then moves onto the next one. Once the entire row is done being scanned, it shifts upwards on the cartesian plane, onto the next row, scanning the window immediately above it. This systematic approach minimizes redundant movement and results in significant reduction in flight time. The scanning pattern of the vehicle is illustrated in Fig. 1 and detailed in Algorithm 1. Let r denote current row number (ranging from 0 to $b_{\max}$) and c denote current column number (ranging from 0 to $l_{\max}$). Let W' be the next window to be scanned.

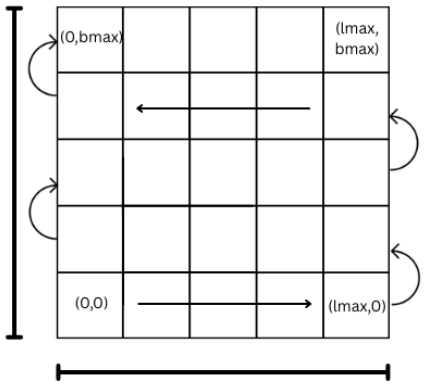


Fig. 1. Working of Sliding Window

Algorithm 1 Grid-Based Sliding Window

Start at grid position $(0, 0)$ and scan to check if the spot is safe or unsafe
if $r \bmod 2 = 0$ **then**
  Move left to right: $W' = (c + 1, r)$
  if $c = l_{\max}$ **then**
    Move vertically: $W' = (l_{\max}, r + 1)$
  end if
else
  Move right to left: $W' = (c - 1, r)$
  if $c = 0$ **then**
    Move vertically: $W' = (0, r + 1)$
  end if
end if
if $b_{\max} \bmod 2 = 0$ **then**
  Stop scanning at $(l_{\max}, b_{\max})$
else
  Stop scanning at $(0, b_{\max})$
end if

C. Flood-fill Algorithm

Flood-fill has been implemented to ensure comprehensive coverage of the terrain for the identification of necessary safe spots. In scenarios where a safe spot partially appears within the current sliding window, it is possible that the remaining portion lies in adjacent windows. Since the sliding window algorithm processes the environment sequentially along a fixed linear path, such partial detections may be overlooked. In order to avoid this, if part of the current window is found to be safe, the algorithm triggers scanning of neighboring windows. This ensures that overlapping safe spots are fully captured. By combining the sliding window algorithm with flood fill, the ability of the system to identify safe spots is significantly improved, making for a more efficient solution.

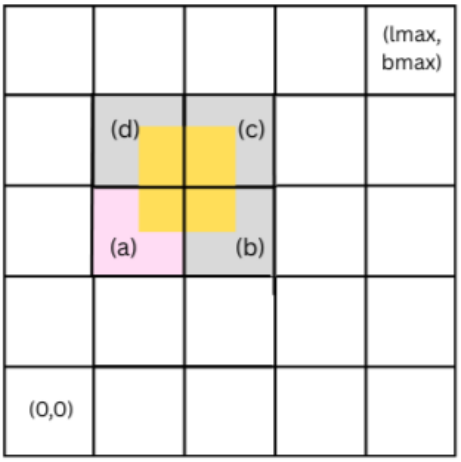


Fig. 2. Working of Flood-fill

Considering Fig. 2, the grid labeled (a) represents the current window. A portion of a safe spot, indicated by the yellow shading, is detected within this window. To ensure the complete region is identified, the flood-fill algorithm consequently scans the adjacent windows (b), (c) and (d) to uncover the overlapping safe spot (represented by the yellow square).

D. Ranking System

When the drone has to make an emergency landing due to unforeseen circumstances, it is crucial for it to know the nearest viable landing spot that minimizes the risk of further damage. This is where a ranking system comes in. If the safe spot that we identified most recently is too far from the drone's current position, it is necessary to select an alternative spot that is closer in proximity, despite having a slightly higher level of risk. When a safe or partially safe region is detected, it may be added to memory and ranked according to a safety metric. If the highest priority slot is reserved for the completely safe spot where landing poses minimum threat, the subsequent slots may be for other spots where landing can pose moderate risk but ends up being a necessity during emergencies. This hierarchical system ensures that, even in the

absence of an ideal landing site, the drone can still execute a controlled descent to the safest available location nearby. This is especially useful in emergency scenarios which may occur due to low battery, overheating or system failure.

V. Hardware Settings

The drone is built on a carbon-fiber quadcopter frame with a 500 mm wheelbase, optimized for strength and low weight. It employs four brushless DC motors (T motors) rated at 920 KV, each paired with 10-inch two-bladed, self-locking propellers made of glass fiber. The motors are controlled by 20A electronic speed controllers compatible with 3S-4S LiPo batteries.Navigation and perception are handled through a stereo vision RGB-D camera with a horizontal field of view of 87°, and a 2D scanning LiDAR with a 180° field of view and range of up to 100 meters. An optical flow sensor and the Pixhawk 5x in-built altimeter further support precise altitude stabilization and terrain-aware landing.The avionics subsystem is centered on the Pixhawk 5x flight controller equipped with a 32-bit ARM processor, which processes data from an inertial measurement unit, magnetometer, and barometer. Communication is managed through a 433 MHz telemetry module supporting full-duplex data transmission. Power is supplied by a 14.8 V, 6200 mAh 4S lithium polymer battery, regulated via a high-voltage power module and distribution board. All components are unified through a control architecture that enables fully autonomous operations, including takeoff, hovering, navigation, and precision landing, with additional provisions for fault-tolerant behavior and emergency response in the absence of GPS or external localization aids.

Fig. 3. Drone Top View

VI. Results & Discussions

The primary objective of this project was to develop and validate a GPS-independent navigation framework for unmanned aerial vehicles (UAVs). As an initial step toward full autonomy, the system was evaluated for its ability to maintain stable hovering without relying on satellite-based positioning systems. These trials were conducted in a controlled outdoor environment, intentionally simulating a GPS-denied scenario.

The drone was deployed in an open test field where GPS signals were deliberately disabled. Relying solely on onboard sensors, including a stereo depth camera, LiDAR, and an optical flow sensor the UAV successfully achieved stable hovering. The tests confirmed the feasibility of sensor-based localization and control in the absence of GPS. The drone maintained its position with minimal drift and exhibited consistent altitude hold and lateral stability over extended durations. These results highlight the effectiveness of the sensor fusion framework in interpreting environmental data and translating it into real-time control responses. Specifically, the combination of stereo depth and LiDAR enabled accurate altitude estimation and obstacle awareness, while the optical flow sensor significantly contributed to lateral stabilization, especially in low-altitude or low-texture environments.

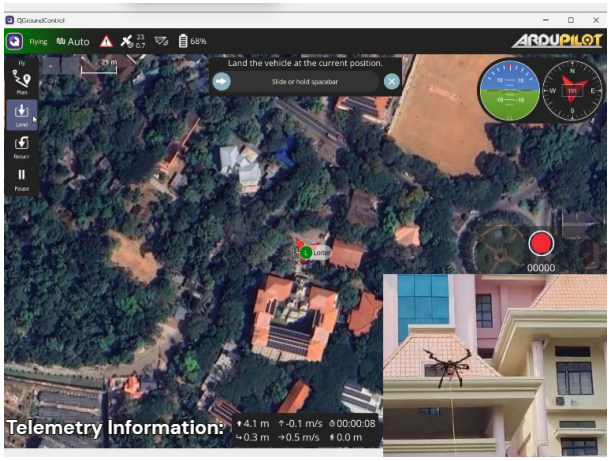


Fig. 4. Drone Flight

Having established reliable GPS-free hovering, the project now advances toward the integration of established autonomous navigation algorithms namely, the Sliding Window Localization Algorithm and the Tunnel Vision Path Planning Strategy. These algorithms are expected to extend the drone's functionality from static hovering to dynamic navigation in structured and constrained GPS-denied environments. Although both algorithms are well-documented in literature, their real-time deployment within the current sensor configuration is still under development. Initial integration steps, including sensor synchronization, coordinate frame transformations, and environmental data collection, have been completed, laying the groundwork for real-time implementation.

Future work will focus on completing the real-time implementation and optimization of the sliding window and tunnel vision algorithms. Once integrated, the system will be tested in semi-structured and cluttered GPS-denied environments to evaluate its performance in real-world scenarios. Quantitative assessments will be carried out to measure localization accuracy, drift rates, trajectory tracking performance, and system robustness under dynamic conditions.

VII. Conclusion

In this work, we present an autonomous navigation system designed for drones operating in GPS-denied environments. Our approach integrates a stereo depth camera, LiDAR and an optical flow sensor with a collection of algorithms - tunnel vision, sliding window, flood-fill and a ranking system for the landing spots. Together, these components can enable efficient terrain analysis as well as identification of safe landing spots, especially in emergency scenarios. The proposed system demonstrates potential for use in real-world applications such as disaster response, traversal through hazardous areas, military defense and extraterrestrial exploration, where traditional navigation algorithms may fall short. While the current implementation is under development, ongoing efforts are focused on further refining and optimizing the system. Future work will focus on perfecting the algorithms and on expanding the dataset for training the safe spot classifier.

References

[1] S. Saha, S. Roy, and S. Saha, "Deep learning in unmanned aerial vehicles: A review," AI Open, vol. 5, pp. 118–136, 2024.
[2] M. H. Bakillah, A. H. Almadani, and B. B. Gupta, "Recent developments in the use of unmanned aerial vehicles (UAVs) for disaster monitoring and management," Computers, Environment and Urban Systems, vol. 99, p. 101998, 2023.
[3] CFD Flow Engineering, "Working principle and components of drone," 2023.
[4] Instructables, "Autonomous Drone," 2021.
[5] A. V. Patel et al., "Risk aware planning with energy efficiency for Martian lava tube exploration using UAV," Heliyon, vol. 10, no. 2, p. e26351, 2024.
[6] M. Radoglou-Grammatikis, P. Sarigiannidis, I. Moscholios, and M. Lagos, "A compilation of UAV applications for precision agriculture and smart farming," IEEE Access, vol. 7, pp. 140088–140112, 2019.
[7] NASA, "Propagation Issues for Communication Between Earth and Mars," MarsPub Technical Paper, TP-2000-209756, Sec. 7, 2000.
[8] A. V. R. Katkuri et al., Autonomous UAV Navigation Using Deep Learning-Based Computer Vision Frameworks: A Systematic Literature Review, PDF, 2024.
[9] A. Patel, S. Karlsson, B. Lindqvist, C. Kanellakis, A. Agha-Mohammadi, and G. Nikolakopoulos, "Towards energy efficient autonomous exploration of Mars lava tube with a Martian coaxial quadrotor," Advances in Space Research, vol. 71, no. 11, pp. 3837–3854, 2023.
[10] M. Radotich, S. Withrow-Maser, Z. deSouza, S. Gelhar, and H. Gallagher, "A study of past, present, and future Mars rotorcraft," in Proc. Vertical Flight Society's 9th Biennial Autonomous VTOL Technical Meeting, NASA Ames Research Center, Jan. 2021.
[11] X. Ézara, G. Rodríguez, J. Seves, R. Rebolo, and M. Alazraki, "AERIAL, an ESA drone approach to conquer Mars atmosphere," ESA GNC Conference Paper, 2022.

SMART SENTRY: UNAUTHORIZED PARKING MONITORING WITH OWNER ALERTS VIA API INTEGRATION

Spoorthi P A
Department of Electronics and Communication Engineering
Dr. Ambedkar Institute of Technology
Bengaluru -560056, Karnataka, India
spoorthiyadav.24@gmail.com

Mala Swadi
Department of Electronics and Communication Engineering
Dr. Ambedkar Institute of Technology
Bengaluru -560056, Karnataka, India
malasinnoor.ec@drait.edu.in

Madhu Shree S
Department of Electronics and Communication Engineering
Dr. Ambedkar Institute of Technology
Bengaluru -560056, Karnataka, India
madhushree2801@gmail.com

Abstract——In the realm of parking enforcement, the deployment of a significant workforce to identify unauthorized parking and impose fines has become common practice. However, this approach is marred by challenges such as bribery and intimidation, which allow vehicle owners to elude penalties. Additionally, the manual search for illegally parked vehicles by towing vans incurs substantial expenses encompassing personnel remuneration, fuel, and physical surveillance. In response, we present an innovative solution aimed at automated detection of unauthorized parking incidents, coupled with real-time alerts. Our proposed system entails the incorporation of Radio Frequency Identification (RFID) tags in all vehicles. An RFID receiver circuit is strategically positioned in zones designated for authorized parking. When a vehicle occupies such a space, its RFID tag engages with the nearby receiver circuit. Upon activation, the RFID reader captures the transmitter ID, promptly notifying the vehicle owner via WhatsApp. This instantaneous alert empowers the owner to swiftly address the unauthorized parking situation. By leveraging RFID technology, this system minimizes the need for extensive human intervention, thus curtailing overhead costs linked to personnel, fuel, and surveillance efforts. This abstract encapsulates an innovative approach to parking management, offering heightened efficiency in unauthorized parking detection and response.

Keywords—***"IoT Implementation", "Vehicle Authorization", "Remote Monitoring", "Intelligent Parking System", "RFID Authorization"***

I. INTRODUCTION

While community living has its own set of apparent advantages, with the good comes the bad. There are certain small issues that can sometimes be quite irritating when it comes to living together in a community. One such common issue that is often a talking point between members of the housing community is parking and the woes it can generate. With more and more people investing in cars, the problem of car parking has also increased.

Unauthorized parking refers to the act of parking a vehicle in a location that is not designated or permitted for parking, such as no- parking zones, fire lanes, pedestrian walkways, or private property without permission. This problem arises due to a combination of factors, including the lack of parking infrastructure, limited awareness for parking regulations, and the sheer volume of vehicles on the roads.

One of the primary causes of unauthorized parking is the inadequate availability of parking spaces. Rapid urbanization and population growth in cities have outpaced the development of parking infrastructure. As a result, there is often a shortage of designated parking areas to accommodate the increasing number of vehicles. This scarcity of parking spots forces some drivers to park in prohibited areas, causing congestion, inconvenience to others, and potential safety hazards. Another contributing factor is the lack of awareness or disregard for parking regulations among vehicle owners.

Addressing the issue of unauthorized parking requires a multifaceted approach. It involves creating sufficient parking infrastructure to meet the growing demand, implementing clear and comprehensive parking regulations, and increasing public awareness of parking rules and their enforcement. Effective enforcement mechanisms, such as the use of technology for automatic detection and monitoring, can help deter unauthorized parking and ensure compliance.

II. LITERATURE REVIEW

The "Automatic Smart Parking System using Internet of Things (IOT)" published in 2015 offers a simple, economical, and effective solution for managing and mapping parking slots remotely through a web browser. However, the initial setup cost and ongoing maintenance expenses could be substantial, making it less feasible for smaller parking facilities or areas with limited financial resources.

The "Automatic Unauthorized Parking Detector with SMS Notification," published in 2019, aims to identify and fine unauthorized parkers in authorized areas using RFID technology, thus discouraging illegal parking, bribery, and corruption. This RFID-based system provides efficient alert mechanisms through SMS notifications to vehicle owners, but the initial setup cost, dependence on cellular networks, and privacy concerns are limitations to consider.

The "Automated Parking System With Bluetooth Access," published in 2014, has the advantage of being cost-effective as the Bluetooth reader and device are cheaper compared to NFC and other readers, and Bluetooth is available in every mobile phone. However, the installation period and initial setup cost are high, and the Lastly, the

"Smart Parking System Using the Raspberry Pi and Android," published in 2017, enables users to efficiently find vacant parking spots through an Android app, enhancing parking management and utilization by employing the Raspberry Pi as a scalable central processing unit for widespread deployment. Despite these benefits, user adoption could be hindered by the need to download and use a specific Android app, which may discourage some drivers from using the system.

In conclusion, while each of these smart parking solutions presents unique advantages in terms of cost efficiency, technological integration, and user convenience, they also face significant challenges related to setup costs, maintenance, network dependency, and user adoption. We aim to address these limitations by reducing setup costs, enhancing user privacy, and simplifying the user interface to encourage widespread adoption. Integrating emerging technologies such as enhanced connectivity further optimizes the efficiency and reliability of these systems, ultimately providing a more seamless parking experience for users.

III. OBJECTIVES

1. Improve Efficiency: The system aims to enhance the efficiency of unauthorized parking enforcement by automating the detection process.
2. Reduce costs: The proposed system aims to minimize overhead costs associated with manpower payment, fuel consumption, and physical surveillance.
3. Enhance enforcement effectiveness: The system intends to improve the effectiveness of enforcing fines by reducing the opportunities for owners to evade penalties through illegal means.
4. Improve owner awareness: The system includes sending notifications to the parking space owners, informing them about the unauthorized parking in their space.
5. Streamline operations: The proposed system aims to streamline parking enforcement operations by providing real-time data and alerts to the owner/authorities.

IV. METHODOLOGY

The devised parking space management system boasts a sophisticated operational approach aimed at seamlessly managing parking occupancy and enhancing communication between the system and the owner. This methodology leverages cutting-edge technology, including Wi-Fi connectivity, RFID integration, ultrasonic sensing, and WhatsApp API for a holistic and efficient parking management experience.

Wi-Fi Connection Establishment and Notification: Upon activation, the system promptly establishes a secure Wi-Fi connection. Once connected, the system initiates a notification to the owner via the widely-used and privacy-centric WhatsApp API. This notification confirms the successful Wi-Fi connection establishment and signals the readiness of the RFID device for operation.

Ultrasonic Sensing for Detection: The heart of the system lies in its ultrasonic sensor, which employs precise distance calculations to detect the presence of vehicles or objects within the designated parking space. Upon detection, the system executes a dual-check mechanism: after an initial detection, a recheck occurs after ten seconds to ascertain whether the detected object is stationary or in motion.

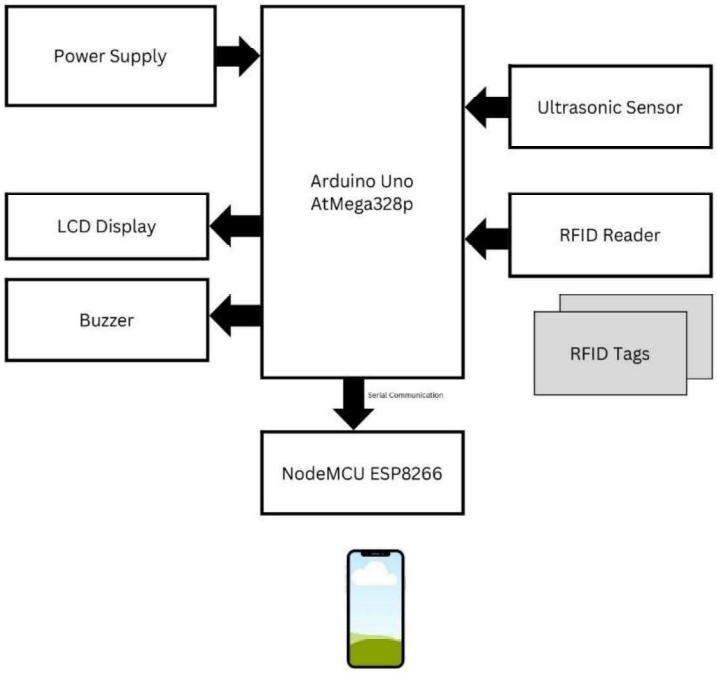


Fig. 1. Block Diagram of Smart Sentry

RFID Integration and Authorization: In cases where the detected object is verified as stationary, the system proceeds with RFID reader activation. This step is pivotal in distinguishing between authorized and unauthorized vehicles. Three distinct scenarios emerge:

A. Unauthorized Object Presence: Should the detected object lack a valid RFID, the system promptly notifies the parking space owner through WhatsApp. This alert underscores the unauthorized presence and prompts appropriate action.

B. Authorized Vehicle with Valid RFID: In instances where a valid RFID is detected, the system authorizes the vehicle, promptly updating the owner with the specific vehicle's identity via RFID tag information. The system then enters a temporary dormancy, later revisiting the authorization status within a predetermined interval.

C. Unauthorized Vehicle with Invalid RFID: When an unauthorized vehicle is detected, but with an invalid RFID, the system promptly informs the owner through WhatsApp. Furthermore, the system ensures periodic reminders every 2 minutes to effectively address the unauthorized presence.

Display and Communication: All pertinent information is elegantly displayed on the LCD display within the parking space, providing transparency and clarity to both vehicle owners and the parking space owner. This visual feedback reinforces the system's reliability and accountability.

Integrated Communication: These intricate processes are facilitated through seamless communication channels established between the system's array of sensors, the microcontroller, and the Wi-Fi module. This robust integration ensures real-time data exchange and cohesive operation of the entire system.

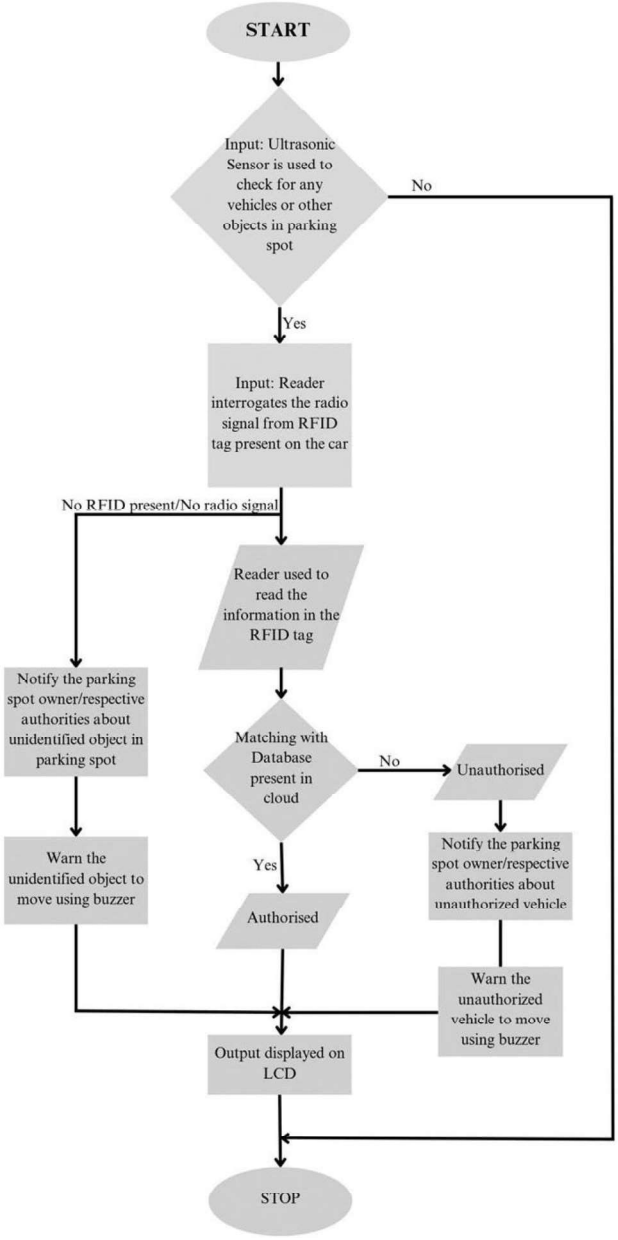


Fig. 2. Flow Chart of Smart Sentry

This comprehensive parking space management system represents a marriage of cutting-edge technology and user-centric design. Its ability to intelligently detect, identify, and communicate with vehicles, combined with its use of secure and efficient communication mediums like Wi-Fi and WhatsApp API, positions it as a model solution for contemporary parking management challenges.

V. RESULTS

Upon activation, the system initiates a seamless experience by visually and digitally confirming successful Wi-Fi connection through intuitive LED indicators and real-time WhatsApp notifications. This visual and digital reassurance establishes a strong foundation for user engagement.

With the Wi-Fi connection solidified, the system engages its core functionality: parking space status monitoring. Users are promptly informed whether the parking space is vacant or occupied. This real- time update serves as a quick reference for users seeking convenient parking.

For occupied spaces, the system seamlessly transitions to the next step: authorization verification. Employing RFID technology, the system cross-checks the presence of valid RFID tags to ascertain the legitimacy of the parked vehicle. This process is designed to ensure security and parking space integrity. Smart Sentry Device instantaneously illuminates its readiness on the LCD display, affirming its operational status to users.

In the absence of any vehicle in the parking space, the LCD display adeptly conveys its status by illuminating with the reassuring message "Ready for Parking."

When an authorized vehicle enters the parking space, the LCD display promptly showcases a welcoming message, confirming the legitimacy of the vehicle's presence. This visual affirmation enhances user experience and provides immediate assurance of a successful entry.

When an unauthorized vehicle or object enters the designated parking space, the system swiftly detects this intrusion through its advanced sensor network. It then initiates a series of immediate responses, including activating the RFID reader and triggering an audible alert, ensuring swift action to address the unauthorized presence.

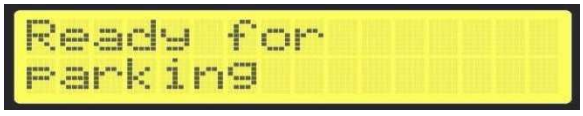
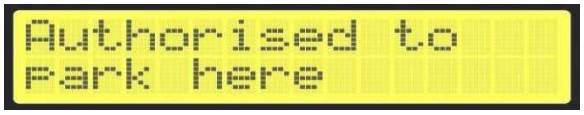
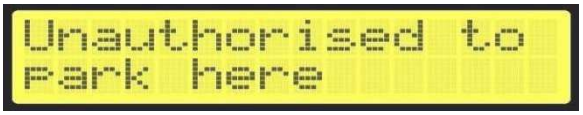
Integral to user satisfaction is real-time information dissemination. The system takes full advantage of the WhatsApp API to provide users with instant updates on parking space occupancy, offering a dynamic and responsive communication channel.

Multiple Vehicle Authorization Made Simple: Addressing users with multiple vehicles, the system offers an elegant solution. By storing authorized RFID tag data for each vehicle, users are informed not only about occupancy but also about which specific vehicle occupies the space. This tailored insight enhances convenience and streamlines decision-making.

Efficient Handling of Unauthorized Occupancy: The inclusion of a strategically placed buzzer proves valuable in addressing unauthorized occupancy. This audible alert discreetly informs unauthorized vehicles to vacate the space promptly, fostering compliance and promoting fair usage.

Continual Monitoring for Real-Time Insights: The system's commitment to real-time data is further reinforced by periodic checks on parking space status. By continually monitoring the space, users receive timely information and can make informed decisions even on the go.

The following images represent output displayed on the LCD for various circumstances:

Cases	Conditions	Message on the LCD display
1	In the absence of any vehicle in the parking space	
2	When an authorized vehicle enters the parking space	
3	When an unauthorized vehicle or object enters the designated parking space	
4	System proactively ensures parking space availability	

VI. CONCLUSION

In the realm of modern living, the Smart Sentry stands as a beacon of innovation. By harmoniously integrating advanced components and user-centric design, it redefines how we perceive and manage parking spaces. Through LED cues and instant WhatsApp alerts, the system ensures users begin their experience with confidence. Real- time updates on parking status, coupled with RFID authorization, elevate security and efficiency. The personalized touch of vehicle- specific notifications caters to individual needs, while the audible buzzer tactfully enforces fair usage. Through this intelligent fusion of technology, the system paves the way for smarter, more convenient parking management, exemplifying the transformative power of innovation in our daily life.

VII. FUTURE SCOPE

The Smart Sentry serves as a foundation for future advancements in parking management and smart urban solutions. This innovative system opens doors to exciting possibilities:

Predictive Analytics: Integrating AI-driven predictive analytics can anticipate parking demand patterns, optimizing space availability and improving user experience.

Smart Payment Solutions: Future iterations might incorporate automated payment systems, eliminating the need for manual transactions and enhancing user convenience.

Renewable Energy Integration: By harnessing renewable energy sources, the system could operate sustainably, reducing its environmental footprint.

Multi-Level Parking Management: Scaling the system to manage multi-level parking structures would provide comprehensive parking solutions for diverse urban environments.

Peer-to-Peer Parking Sharing: Enabling users to share parking spaces among themselves could lead to efficient space utilization and encourage a sense of community.

Smart Navigation Integration: Integrating with navigation systems could guide users to available parking spaces, reducing traffic congestion.

Machine Learning for Security: Utilizing machine learning algorithms could enhance security by detecting anomalous behaviours and patterns in parking areas.

The future scope of the Smart Sentry extends beyond the present capabilities, encompassing a realm where technology continues to reshape urban experiences. By embracing these future possibilities, the system has the potential to revolutionize how we perceive, access, and utilize parking spaces in our ever-evolving cities.

REFERENCES

[1].“Automatic Smart Parking System using Internet of Things (IOT)” by Basavaraju S R RV College of Engineering, International Journal of Scientific and Research Publications ISSN 2250-3153 Volume 5, Issue 12, December 2015

[2].“Automatic Unauthorized Parking Detector with SMS Notification” by Madhurima Chakrabarti , Sandip Karmakar , Sumit Singh , Madhura Sur , Dipankar Kundu International Advanced Research Journal in Science, Engineering and Technology ISSN 2393-8021 Volume 6, Issue 5, May 2019

[3].“Automated Parking System With Bluetooth Access” by Harmeet Singh, Chetan Anand, Vinay Kumar, Ankit Sharma ,International Journal Of Engineering And Computer Science ISSN 2319-7242 Volume 3, Issue 5, May 2014

[4]. “Smart Parking System Using the Raspberry Pi and Android” by Prof. Ashwini Gavali , Pooja Kunnure , Supriya Jadhav , Tejashri Tate , varsha patil International Journal of Computer Science and Information Technology Research ISSN 2348-1196 Volume 5,Issue 2,June 2017

[5].“Automatic vehicle detection and identification using visual features (2017)” by Hao Lyu, University of Windsor, Windsor, Ontario, Canada, MS 2017

[6]. “Car Detection In Live Video Incorporated With Machine Intelligence” by Bechra Nikita, Prof. A. R. KaziInternational Journal of Advance Research and Innovative Ideas in Education, Volume 3, Issue 6, pp. 495-506, 201

Automated Vehicle Registration Number Plate Recognition System using CNN

Mala Swadi
Department of Electronics and Communication Engineering
Dr. Ambedkar Institute of Technology
Bengaluru -560056, Karnataka, India
malasinnoor.ec@drait.edu.in

Spoorthi P A
Department of Electronics and Communication Engineering
Dr. Ambedkar Institute of Technology
Bengaluru -560056, Karnataka, India
spoorthiyadav.24@gmail.com

Aditya
Department of Electronics and Communication Engineering
Dr. Ambedkar Institute of Technology
Bengaluru -560056, Karnataka, India
adityaadi1718@gmail.com

Chaithrashree B S
Department of Electronics and Communication Engineering
Dr. Ambedkar Institute of Technology
Bengaluru -560056, Karnataka, India
chaithrashree.b.s.2004@gmail.com

***Abstract*— The manual identification and monitoring of vehicle registration number plates take a lot of time, are prone to errors, and do not work well, especially in high traffic situations, with various plate designs, and in tough environmental conditions. Traditional automated solutions often face problems like poor image quality, different lighting conditions, and trouble distinguishing number plates from messy backgrounds. These issues affect the accuracy and growth potential of current systems. Therefore, there is a need for a strong, efficient, and automated vehicle number plate recognition system that can tackle these challenges. The main goal is to create a CNN-based Automated Vehicle Number Plate Detection System that can accurately detect, extract, and recognize vehicle registration numbers from images or video streams. This system aims to meet the needs of intelligent transportation systems and support seamless integration with applications like traffic law enforcement, toll collection, parking management, and security monitoring, helping to develop smarter and more efficient urban mobility solutions.**

Key words: Convolutional Neural Networks (CNNs), YOLO, Optical Character Recognition (OCR)

I. INTRODUCTION

The programmed number plate recognition plays an important role in various real-world applications, such as monitoring street traffic, controlling access at parking areas, and managing automated toll collection booths. As the number of vehicles grows daily, keeping track of them manually becomes difficult, which is why we need the VRNPR system. Additionally, the rise in road accidents and traffic congestion also drives the need for the VRNPR system. Traffic systems in developed cities like Dubai, Canada, and Italy are much better compared to developing countries like India, largely due to automated number plate recognition systems.

These systems help ensure that citizens follow traffic rules, as violations can lead to consequences. They have also contributed to a reduction in the number of accidents. Therefore, implementing the VRNPR system across India and other developing countries is essential for progress and citizen safety. VRNPR identifies a vehicle's registration number plate from images captured by cameras. This process combines several techniques, including object detection, image processing, and pattern recognition [1].

VRNPR systems use various methods, such as artificial neural networks, optical character recognition, probabilistic neural networks, back propagation neural networks, inductive learning, and convolutional neural networks [2]. The key task is to detect and recognize the number plate, which is achieved using convolutional neural networks (CNN), along with other techniques. CNNs are chosen for their high accuracy, around 90%, even with relatively small training sizes. Convolutional neural networks are regarded as one of the most effective advancements in computer vision have been considered as one of the most persuasive developments in the field of computer vision.

In recent years, Convolutional Neural Networks (CNNs), a part of deep learning algorithms, have shown outstanding results in image recognition and classification tasks. Using CNNs, an automated vehicle registration number plate recognition system can achieve strong and accurate identification in different environmental conditions like poor lighting, occlusion, or complicated backgrounds [3-4].

[5-6] This project introduces a CNN-based Automated Vehicle Registration Number Plate Recognition System that aims to simplify vehicle identification processes. By taking advantage of CNN architectures, this system seeks to provide high accuracy in extracting and recognizing registration numbers from images or video frames [7-8]. The implementation is designed for real-time applications and also tackles issues such as plate variations across regions, image noise, and different plate orientations [9].

II. LITEARTURE SURVEY

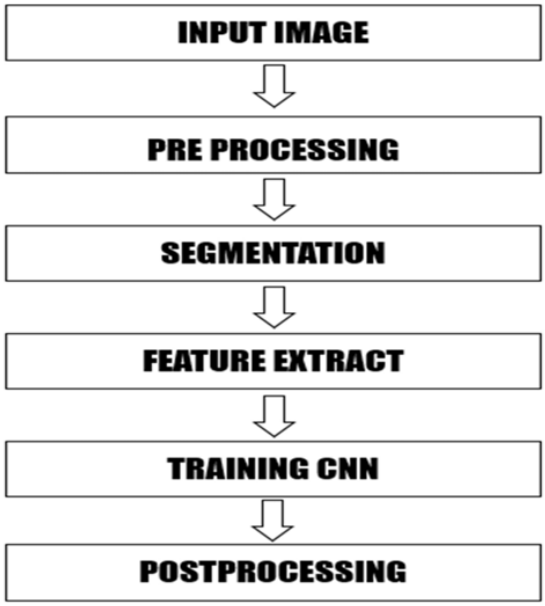
VRNPR, or Vehicle Registration Number Plate Recognition, is basically a system that can automatically detect and read license plates from vehicles. It's really useful for things like tracking traffic, catching speeders, or identifying vehicle owners. The tricky part is that accuracy depends on factors like how fast the car is moving, whether it's day or night, and how clear the image is. To tackle this, newer systems rely on Convolutional Neural Networks (CNNs), which are smart algorithms that recognize patterns in images and can deliver very high accuracy even with limited training data. Researchers compare different VRNPR methods based on

their speed, accuracy, strengths, and weaknesses to make the system more reliable in real-world situations.

The goal of VRNPR (Vehicle Registration Number Plate Recognition) is to automatically read vehicle number plates, which helps in traffic control and identifying rule violators like speeders. While many systems exist, they face challenges such as fast-moving vehicles, poor lighting, and low image quality. To improve accuracy, modern methods use Convolutional Neural Networks (CNNs), which can recognize plates with around 90% accuracy even with limited data. This study reviews different VRNPR techniques, comparing their strengths, weaknesses, speed, and accuracy to build more reliable systems.

In India, all vehicle owners are required to use High-Security Registration Plates (HSRP), as they are tamper-proof, theft-resistant, and easier to identify than old number plates. This step helps standardize plates, making it simpler for authorities to track vehicles and improve road safety—an urgent need since India records the world's highest road accident deaths. To support this, a new model was developed to detect and read HSRPs. A dataset of 500 plate images was created, annotated, and tested using YOLOv5 for plate detection and EasyOCR for character recognition. The model achieved 100% accuracy in plate detection and 80% accuracy in character recognition, showing strong potential compared to existing methods.
Object recognition technology is now widely used to improve security, efficiency, and daily life. One important application is in vehicle recognition systems that help identify and track vehicles for security or service purposes. This paper presents a vehicle route tracking system that uses surveillance cameras to recognize license plates and monitor the movement of a targeted car. By combining template matching for plate recognition with GPS data, the system can track the car's path and current location. It achieves about 80% accuracy in detecting vehicle plates and maps the vehicle's route effectively over time.

III. METHODOLOGY



A flow of image processing.

A. Input Image

The process starts by capturing an image of a vehicle—either in color or grayscale—which contains the number plate. This image serves as the base input for further steps. A Convolutional Neural Network (CNN) then analyzes the image, extracting features and learning patterns to accurately recognize the number plate.

B. Preprocessing

Preprocessing is an important step to prepare images before they are sent to the CNN. It makes the input cleaner, consistent, and easier to process, which improves system performance. Common steps include resizing images to the same dimensions, normalizing pixel values to a standard scale, converting colored images to grayscale for simpler computation, and cropping to focus only on the number plate. Overall, preprocessing ensures the data is optimized for accurate feature extraction and recognition.

C. Segmentation

Segmentation is the process of breaking an image into useful parts. In vehicle number plate recognition, it means first isolating the number plate area (region of interest) and then splitting it into individual characters or digits. This helps the system focus only on the important details. Techniques like bounding boxes (to mark where the plate or characters are) and thresholding (to separate characters from the background) make recognition easier for the CNN.

D. Feature Extract

After segmentation, the CNN analyzes the input image and extracts key features. Through its layers, it first detects simple patterns like edges and curves, then learns complex shapes like numbers and letters. Convolutional layers capture features, pooling layers simplify them to save computation, and feature maps highlight important regions of the image. These extracted features are then passed on for classification and recognition of the number plate.

E. Preprocessing

Training a CNN is essential for building an accurate recognition system. It starts with feeding the model labeled images of number plates, then measuring errors using a loss function. Through optimization, the model adjusts its parameters to reduce these errors. By running several training cycles (epochs), the network gradually learns patterns and becomes capable of accurately predicting number plate characters.

F. Segmentation

The raw output from a CNN often needs cleanup before use. Post-processing refines the results by converting predictions into readable text (like "ABC1234"), correcting minor errors using format rules or dictionaries, and integrating the final recognized plate number with systems such as traffic databases or law enforcement tools. This ensures accurate and usable results, even if the CNN makes small mistakes.

G. Post processing

Once the segmentation is done, the CNN begins to analyze the input and extract relevant features. Feature extraction involves multiple convolutional layers in the network. These layers progressively learn hierarchical patterns, starting from

simple edges and corners to more complex structures like alphanumeric characters.

IV. CONVOLUTIONAL NEURAL NETWORK (CNN)

A Convolutional Neural Network (CNN) is a specific kind of deep learning algorithm that processes structured data like images and videos. It imitates how the human brain handles visual information. This makes it especially good for tasks like image classification, object detection, and facial recognition.

Key Components of CNN

1. Convolution Layers – Apply filters to the input image to detect patterns like edges, textures, and shapes.
2. Pooling Layers – Reduce the size of feature maps, making the model faster and less affected by small changes in the image.
3. Fully Connected Layers – Combine all extracted features to make the final prediction or classification.

CNNs are powerful because they automatically learn features from raw data, reducing the need for manual effort.

How the Data is Used

1. Training Data

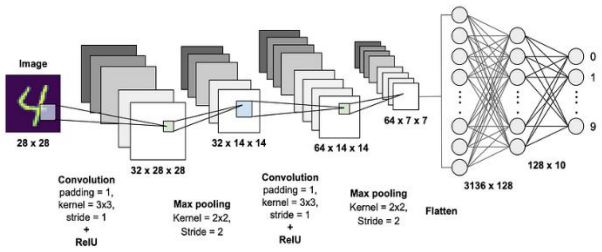
Used to teach the CNN. The network learns patterns by adjusting weights using methods like backpropagation and gradient descent.

Goal: minimize errors (loss) and improve predictions.

2. Testing Data

A separate dataset to check how well the model performs on new, unseen data. Evaluated using metrics like accuracy, precision, recall, and F1 score.

By using separate training and testing data, the CNN prevents overfitting and ensures it works well in real-world situations



VI. IMPLEMENTATION

A. Data Collection and Dataset Preparation

- Images: Gather a diverse dataset of vehicle images containing number plates. The dataset should cover variations in lighting conditions (day/night), Plate formats (regional designs, fonts, and sizes), Camera angles and perspectives, Noise and occlusions (blurred images, glare, or obstructions).
- Labels: Create annotations for:Bounding boxes indicating the location of number plates for detection and Alphanumeric characters for recognition.

B. Preprocessing

- Image Augmentation: Enhance the dataset by generating variations of images (e.g., rotate, scale, brightness adjustments, etc.) to make the model robust.
- Normalization: Normalize pixel values to a fixed range (e.g., [0, 1] or [-1, 1]).
- Image Resizing: Resize all images to a consistent dimension (eg., 333x75 pixels) to standardize input to the CNN.

C. Segmentation (Plate Detection)

- Use a bounding box-based object detection algorithm to locate the region of interest (ROI) i.e., the number plate.
- Model Choices:YOLO (You Only Look Once) and CNN (Convolutional Neural Network)

D. Character Recognition (Using CNN)

Once the plate is segmented, extract and recognize characters from the plate using a CNN based approach.

Steps:

1. Convert the number plate to grayscale.

2. Apply thresholding or edge detection to isolate individual characters.

3. Segment individual characters using contour detection or Connected Component Analysis.

4. Train a CNN for character recognition.

CNN Model for Character Recognition:

- Input: Image of each character (e.g., 28x28 pixels)
- Layers: Convolutional, pooling, and fully connected layers.
- Output: Predicted alphanumeric character.

E. Training the System

- Train the plate detection and character recognition models separately.
- Use labeled datasets with bounding boxes for detection and character images for recognition.
- Loss Functions:

Detection: Binary Cross-Entropy or IoU-based loss.

Recognition: Categorical Cross-Entropy.

- Optimization: Adam or SGD optimizers.
- Training Dataset: Total 864 images belonging to 36 classes.
- Testing Dataset: Total 216 images belonging to 36 classes.

F. Postprocessing

- After character recognition, combine the individual characters into a complete registration number.
- Validate the format of the detected number plate against regional rules (e.g., length, pattern).
- Use error-correction techniques, such as matching against a database of valid registration numbers.

Assuming predictions is a list of characters

Predictions :

Numbers : '0', '1', '2'…… , '8', '9'

Alphabets : 'A', 'B', 'C'…… , 'Y', 'Z'

VII. RESULTS

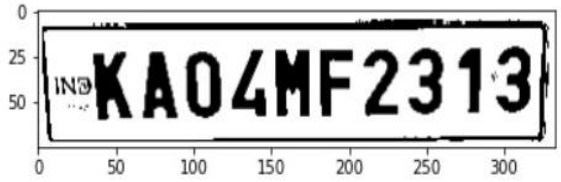
We use image processing where a captured vehicle number plate image is passed through a trained CNN model, which converts the alphanumeric characters into text and displays the result.

A. Vehicle Image Capture :

- Input: Capture an image or video frame containing the vehicle and its number plate using a camera.
- Output: Digital image of the vehicle.

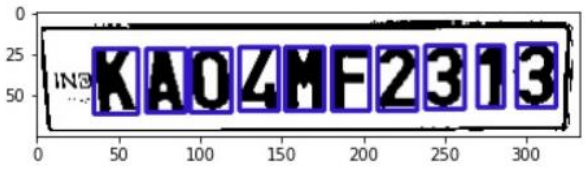
B. Pre-Processing :

- Convert the image to grayscale to reduce complexity while retaining necessary details.



C. Character segmentation :

- Extract individual characters from the binarized license plate image using techniques like connected component analysis.
- Bounding Boxes: The object detection model outputs bounding box coordinates around detected number plates.
- Contour Detection: Identify individual characters using contours.



D. Character Recognition :

- Apply Optical Character Recognition (OCR) to recognize characters and convert them to text.
- We use a CNN or pre-trained OCR model to extract alphanumeric characters.

A. CNN Trained Predicted characters :

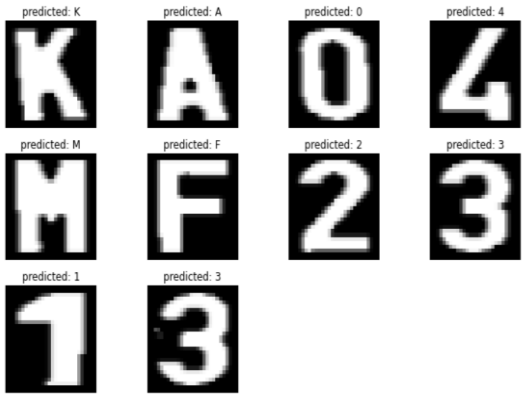
- CNN for OCR: Design or use a pre-trained CNN model (e.g., Tesseract OCR, CRNN) for recognizing alphanumeric characters.
- Training the CNN:

Input: Individual characters extracted from number plates.

Output: Class labels for characters (e.g., 0-9, A-Z).

- Data Augmentation: Augment character images to handle different fonts, sizes, and distortions.

B. Final Output :



VIII. ANALYSIS

TensorFlow is an open-source machine learning framework developed by Google. It's one of the most popular tools for building and deploying machine learning (ML) and deep learning (DL) models. With TensorFlow, you can create different types of neural networks and handle complex numerical computations efficiently.

TensorBoard is a companion visualization tool that comes with TensorFlow. It helps you track and understand how your models are learning by showing metrics, training progress, data flow graphs, and even changes in model weights. This makes it much easier to monitor, debug, and improve your models.

An **epoch** simply means one full pass through the entire training dataset. During each epoch, the model goes through all the training data once, performs forward and backward passes, and updates its weights to improve learning.

If we say **Epochs = N**, it means the model will repeat this process **N times** with the full dataset**.**

Uses of Tenser flow :

- Optimizes computations for performance and memory.
- Enables deployment on mobile, web, and embedded systems.

- Visualizes and debugs models with TensorBoard.

Key Observations
Epoch 1:
- Training Loss: 3.4880
- Training Accuracy: 7.18%
- Validation Loss: 3.3180
- Validation Accuracy: 21.30%

Epoch 2:
- Training Loss: 3.1436
- Training Accuracy: 18.17%
- Validation Loss: 2.8440
- Validation Accuracy: 38.43%

Epoch 3:
- Training Loss: 2.5766
- Training Accuracy: 34.26%
- Validation Loss: 2.2535
- Validation Accuracy: 58.80%

FORMULA TO CALCULATE THE ACCURACY AND LOSS OF TRAINED AND VALIDATION SETS

$$\text{Training Accuracy} = \frac{1}{N}\sum_{i=1}^{N} 1(\hat{y}_i = y_i) \times 100$$

$$\text{Training Loss} = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}(y_i, \hat{y}_i)$$

Where:
- N: Number of samples in the training set
- Yi: True label for the i-th training sample
- y^i: Model's predicted output for the i-th training sample
- L(yi,y^i): Loss function for the i-th sample (e.g., Mean Squared Error, Cross Entropy Loss, etc.)

$$\text{Validation Accuracy} = \frac{1}{M}\sum_{i=1}^{M} 1(\hat{y}_i = y_i) \times 100$$

$$\text{Validation Loss} = \frac{1}{M}\sum_{j=1}^{M} \mathcal{L}(y_j, \hat{y}_j)$$

Where:
- M: Number of samples in the validation set
- yj: True label for the j-th validation sample
- y^j: Model's predicted output for the j-th validation sample
- L(yj,y^j): Loss function for the j-th sample

Epoch 79:
- Training Loss: 0.0852
- Training Accuracy: 97.69%
- Validation Loss: 0.0692
- Validation Accuracy: 98.61%

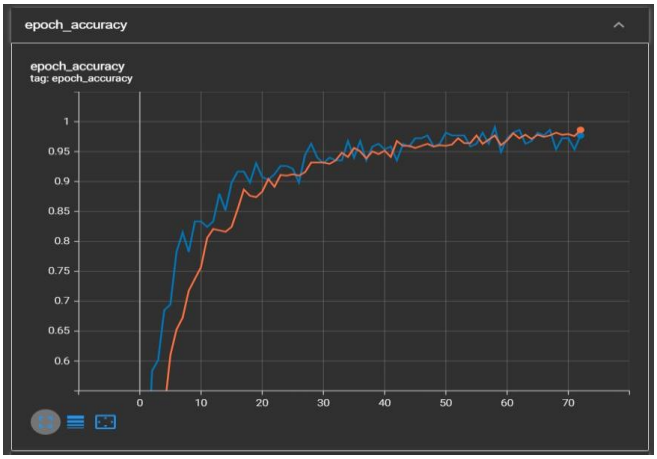
Epoch 80:
- Training Loss: 0.0894
- Training Accuracy: 97.22%
- Validation Loss: 0.1064
- Validation Accuracy: 95.83%

Trained data Observations :
- Peak Performance at Epoch 79 : The validation accuracy reached 98.61%, which is the highest achieved during the training process.
- Consistent High Training Accuracy : The training accuracy remains stable above 97%, showing the model has thoroughly learned the patterns in the training dataset.

How Epochs Affect Accuracy:
- Few Epochs : The model underfits, leading to low accuracy on both training and validation sets.
- Optimal Epochs : The model learns enough patterns to generalize well, maximizing validation accuracy.
- Large Epochs : The model overfits, leading to high accuracy on both training and validation sets.

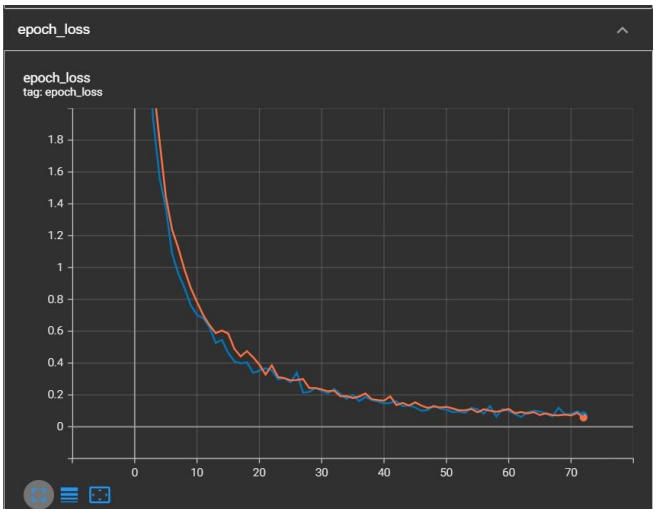
Sl.No	Epoches	Taining Loss	Training Accuracy	Validation Loss	Validation Accuracy
1	10	0.9233	0.7211	0.8105	0.8194
2	20	0.4420	0.8646	0.3539	0.9213
3	30	0.2823	0.9039	0.2254	0.9259
4	40	0.2038	0.9444	0.2228	0.9630
5	50	0.1343	0.9595	0.1153	0.9722

Tensor Board plot of epoch accuracy during training



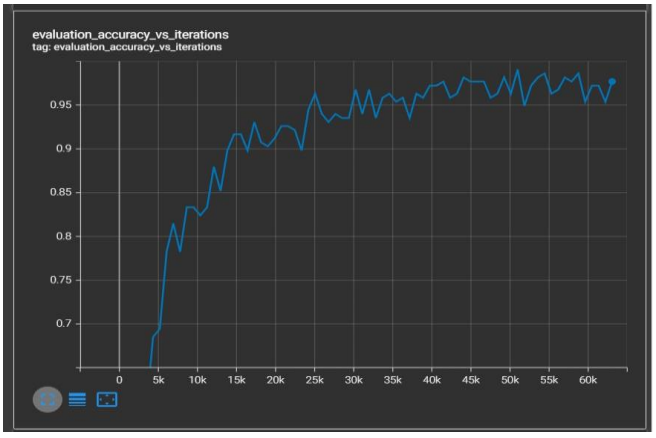
The graph shows that the model quickly learns during the first few epochs, with accuracy rising sharply in the beginning. After about 20 epochs, the progress slows down, and the accuracy levels start to settle. Both training and validation accuracy move closely together, which means the model is not just memorizing but also working well on new data. There are small ups and downs after 30 epochs, but the accuracy stays strong above 90%. By the end, the model reaches around 97–98% accuracy with almost no gap between training and validation, showing it's well-trained and performing reliably.

Tensor Board plot of epoch loss during training

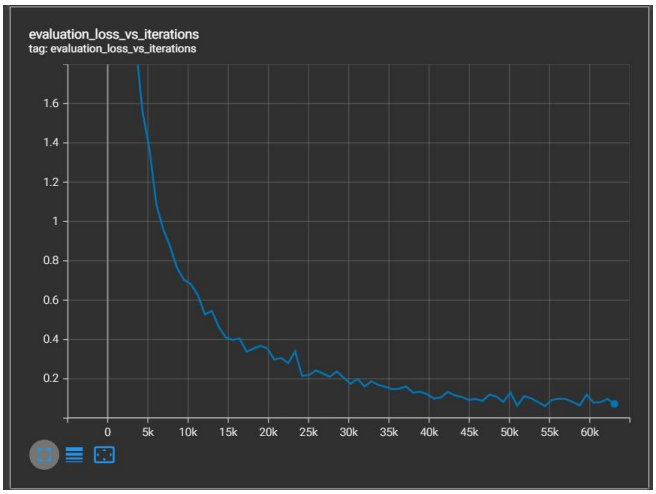


This graph shows how the model's loss decreases steadily as training progresses. At the beginning, the loss is quite high, but it drops quickly within the first few epochs, meaning the model is learning fast in the early stages. After about 20 epochs, the loss continues to decline but at a slower and smoother rate, showing that the model is fine-tuning its learning. Both the training and validation loss curves follow almost the same path, which suggests that the model is not overfitting and is generalizing well to new data. By the end of training, the loss values are very close to zero, indicating that the model has achieved strong performance with minimal errors.

Tensor Board plot of epoch evaluation accuracy vs iterations



Tensor Board plot of epoch evaluation loss vs iterations



CONCLUSION AND FUTURE SCOPE

The Automated Vehicle Number Plate Recognition (AVNPR) system powered by Convolutional Neural Networks (CNNs) is a smart solution for modern traffic and security needs. CNNs can automatically learn features from images, making them highly accurate in detecting and recognizing number plates under different conditions like poor lighting, angles, or varying plate designs. This makes the system useful in traffic management, toll collection, parking, and law enforcement. It also reduces manual effort and speeds up processing. While challenges remain—such as handling damaged or unclear plates and the need for high computing power—CNN-based systems still offer a reliable, efficient, and scalable way to improve vehicle monitoring and overall transportation management.

Future improvements in Automated Number Plate Detection can handle blurry or mirrored images, integrate with IoT for smart city use, support EV charging and emission zone monitoring, and work with AI to predict incidents and optimize routes.

REFERENCES

[1] CNN based Automated Vehicle Registration Number Plate Recognition System . Sachin Shrivastava, Dept. of Computer Sc. & Engg, Sanjeev Kumar, Dept. of I.T., Singh Kapil Shrivastava, Dept. of Computer Sc. & Engg, Vishnu Sharma, Dept. of Computer Sc. & Engg. 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN) 2022 J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] Recognition of High-Security Registration Plate for Indian Vehicles. Sweta Rani, Vivek Shukla, Ramesh Kumar Mohapatra. 3rd International conference on Artificial Intelligence and Signal Processing (AISP) 2023K. Elissa, "Title of paper if known," unpublished.

[3] Vehicle Route Tracking System based on Vehicle Registration Number Recognition using Template Matching Algorithm. Lai Chor Kiew, Abu Jafar Md Muzahid, Syafiq Fauzi Kamarulzaman. International Conference on Software Engineering & Computer Systems ICSECS 2021.

[4] Vehicle Number Plate Detection through live stream using Optical Character Recognition (OCR). Ananya Sri Shetty , V Sai Vineeta , Sreya Ravi , Nerella Likhitha. 7th International Conference on Trends in Electronics and Informatics (ICOEI) 2023.

[5] Saquib Nadeem Hashmi, Kaushtubh Kumar, Siddhant Khandelwal, Dravit Lochan, Sangeeta Mittal,(2019) "Real Time License Plate Recognition from Video Streams using Deep Learning", International

Journal of Information Retrieval Research, Volume 9 • Issue 1 • January-March 2019, IGI Global, Web of Science Emerging Sources Citation Index (ESCI), ISSN: 2155-6377, EISSN: 2155- 385 DOI: 10.4018/IJIRR.

[6] Hendry and Rung-Ching Chen, (2019) “Automatic License Plate Recognition via sliding window darknet-YOLO deep learning”, Image and Vision Computing 87 47 56, Elsevier B.V, Science Citation Index(SCI), doi:10.1016/j.imavis.2019.04.007

[7] Piotr Lubkowski, Dariusz Laskowski,(2017) “Assessment of Quality of Identification of Data in Systems of Automatic Licence Plate Recognition” In: Mikulski J. (eds) Smart Solutions in Today’s Transport. TST 2017. Communications in Computer and Information Science, vol 715. Cham, Springer, doi:10.1007/978-3 319-66251-0_39.

[8] New Ni Kyawf, G R Sinhaf, Khin Lay Mon, (2018) “License Plate Recognition of Myanmar Vehicle Number Plates A Critical Review," IEEE 7th Conference on Consumer Electronics,978-1 5386-6309-7/18/$31.00 ©2018 IEEE.

[9] Abhishek Kashyap, Suresh, Anukul Patil, Saksham Sharma, Ankit Jaiswal,(2018) International “Automatic Number Plate Recognition”, Conference on Advances in Computing, Communication Control and Networking, 978-1-5386-4119 4/18/$31.00 ©2018 IEEE.

[10] Chirag Patel, Dipti Shah, Atul Patel,(2013) “Automatic Number Plate Recognition System (VRNPR): A Survey”, International Journal of Computer Applications.

[11] R. Ghosh, S. Thakre and P. Kumar, "A vehicle number plate recognition system using region-of-interest based filtering method",(2018) Conference on Information and Communication Technology (CICT), Jabalpur, India, 2018, pp. 1-6, doi:10.1109/INFOCOMTECH.2018.8722345.

Defending Machine Learning: GAN-Driven Detection of Adversarial Data Poisoning

Mohammed Nayeem
Trine University
Angola, Indiana, USA
Email: nm2751478@gmail.com

***Abstract*—This paper offers an in-depth exploration of contemporary cybersecurity challenges by dissecting critical attack methodologies and their corresponding mitigation frameworks. We analyze the intricate mechanics of insidious side-channel exploits, the disruptive potential of distributed denial-of-service (DDoS) campaigns, and the pervasive risks posed by cross-site scripting (XSS) vulnerabilities. Furthermore, we delve into specialized reconnaissance techniques and evaluate the capabilities of modern offensive security tools in understanding adversarial tactics. This comprehensive overview aims to equip cybersecurity professionals with enhanced awareness and actionable insights for fortifying digital infrastructures against an evolving array of sophisticated threats.**

I. Introduction and Motivation

Machine learning systems have rapidly evolved into critical components of modern digital infrastructure, powering applications in domains such as healthcare, finance, cybersecurity, and autonomous systems. However, their reliance on large-scale data makes them inherently vulnerable to adversarial manipulations, particularly data poisoning attacks [1]. In these attacks, an adversary subtly injects maliciously crafted data into the training set to corrupt the model's behavior, often in ways that remain imperceptible to traditional validation techniques. Such attacks pose serious risks: from undermining spam filters and fraud detection systems to compromising diagnostic models in medical imaging [2].

The central motivation for this work arises from the inadequacy of conventional anomaly detection approaches, such as Support Vector Machines (SVMs) and clustering-based techniques, which often fail to capture the highly adaptive nature of poisoning strategies [3]. Generative Adversarial Networks (GANs) offer a promising alternative, as they naturally excel at learning complex data distributions and identifying subtle deviations. By harnessing the adversarial interplay between generator and discriminator networks, GAN-based systems can detect poisoned samples as out-of-distribution anomalies, thereby improving resilience against adversarial manipulation.

Another motivation lies in the rising sophistication of poisoning campaigns. Early poisoning attacks often relied on blunt statistical outliers that were easier to detect. Today's attacks, however, are designed to be stealthy—embedding backdoors, mimicking normal data characteristics, or leveraging clean-label strategies to bypass conventional defenses [4]. These realities demand more robust detection mechanisms that can adapt to evolving adversarial tactics.

This paper makes three key contributions. First, it develops a taxonomy of poisoning threats, highlighting the distinct challenges posed by clean-label, backdoor, and targeted poisoning attacks. Second, it proposes a GAN-driven framework for poisoned data detection, where the discriminator is trained to model clean data distributions and reject poisoned deviations. Finally, it provides a comparative evaluation against baselines, including SVMs and traditional GANs, to demonstrate the efficacy and limitations of the proposed approach.

The motivation for this research extends beyond technical curiosity—it addresses a pressing need for trustworthy AI systems in mission-critical environments. From securing financial fraud detection pipelines to safeguarding autonomous vehicle decision-making, the ability to detect and mitigate poisoned training data is essential for ensuring the integrity, fairness, and safety of machine learning deployments in adversarial settings [5], [6].

II. Theoretical Foundations of GAN-Based Anomaly Detection

Generative Adversarial Networks (GANs) were first introduced by Goodfellow et al. as a minimax game between two neural networks: a generator G and a discriminator D [7]. The generator attempts to produce synthetic samples that mimic the distribution of real data, while the discriminator is trained to distinguish between genuine and generated samples. This adversarial process enables GANs to capture highly complex data distributions, making them suitable for anomaly detection tasks where poisoned or adversarial data represents deviations from the expected distribution.

Formally, the GAN training objective is expressed as:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{\text{data}}}[\log D(x)] + \mathbb{E}_{z\sim p_z}[\log(1-D(G(z)))].$$

Here, p_{data} represents the true data distribution, and p_z is the prior distribution for latent variables. The discriminator D aims to maximize its ability to classify real versus generated data, while the generator G attempts to minimize the same objective by producing samples indistinguishable from real ones.

In the context of anomaly detection, the discriminator can be leveraged to act as a probabilistic scoring mechanism. Samples that fall outside the learned distribution receive lower discriminator confidence, signaling potential anomalies such as poisoned data [8]. This makes GANs particularly effective for

adversarial poisoning detection, where the goal is to identify data points that are intentionally crafted to remain stealthy.

Several extensions of GANs have been developed to address challenges such as training instability and mode collapse. For instance, Least Squares GAN (LSGAN) modifies the loss function to provide smoother gradients, thereby stabilizing training and improving the quality of learned distributions [9]. Wasserstein GANs (WGANs) introduce an Earth-Mover (EM) distance-based loss that mitigates vanishing gradients and enhances convergence [10]. These improvements are crucial for anomaly detection tasks, where small deviations must be reliably captured without collapsing the generative process.

Another theoretical foundation relevant to anomaly detection lies in reconstruction-based approaches. By coupling GANs with autoencoder architectures, researchers have designed models that learn compact latent representations of clean data. Anomalies, including poisoned samples, exhibit higher reconstruction errors when passed through these architectures, providing another detection metric [11]. This hybrid strategy combines generative modeling with representation learning to strengthen robustness.

GAN-based anomaly detection aligns well with adversarial defense because it mimics the adversary–defender dynamic. While the generator learns to simulate poisoning attempts, the discriminator continuously adapts to distinguish them, thereby emulating real-world adversarial environments. This co-evolutionary training dynamic ensures that the discriminator remains effective even as poisoning strategies grow more sophisticated [12].

Finally, the theoretical advantage of GANs lies in their ability to model high-dimensional distributions without requiring explicit density estimation. This property is essential for complex data modalities such as images, text, and network traffic, where adversarial perturbations may be imperceptible yet malicious. By framing anomaly detection as a distribution-matching problem, GANs provide a principled approach for detecting poisoned data embedded within large training corpora.

III. Adversarial Data Poisoning: Taxonomy and Threat Models

Adversarial data poisoning is a deliberate manipulation of training datasets with the goal of corrupting the learned model [1], [6]. Unlike traditional noise or random corruption, poisoning attacks are carefully crafted to be both stealthy and impactful. Understanding the taxonomy of these attacks and the underlying threat models is crucial for designing effective defenses.

A. Taxonomy of Poisoning Attacks

Poisoning attacks can be broadly classified into several categories:

- **Availability Attacks:** The goal is to degrade overall model accuracy by flooding the training dataset with corrupted samples, making the model unreliable [3].
- **Integrity (Targeted) Attacks:** These attacks aim to manipulate the model into misclassifying specific instances, while leaving general accuracy largely unaffected. Backdoor attacks are a typical example.
- **Clean-Label Attacks:** Here, the attacker does not modify labels but introduces carefully crafted adversarial samples that blend seamlessly into the dataset. These attacks are particularly difficult to detect [4].
- **Backdoor Attacks:** The adversary implants a hidden trigger pattern in training samples. At inference time, the presence of this trigger forces the model into a specific misclassification [13].

B. Threat Models

Poisoning effectiveness depends on the adversary's knowledge and access:

- **White-Box Model:** The attacker has full access to the training pipeline, including data preprocessing and model parameters.
- **Gray-Box Model:** The attacker has partial knowledge, e.g., access to training data but limited knowledge of model architecture.
- **Black-Box Model:** The attacker can only query the model but has no direct visibility into training details.

These distinctions are critical because defense strategies vary significantly depending on the assumed adversarial knowledge.

C. Attack Pipeline Illustration

Figure 1 illustrates the typical poisoning attack pipeline. The adversary injects poisoned data into the training dataset, which compromises the model during training, leading to biased or malicious outcomes during inference.

IV. Proposed Framework: GAN-Driven Poisoning Detection

To address the limitations of traditional anomaly detection methods, we propose a GAN-driven framework designed specifically for identifying poisoned training data. The central idea is to leverage the discriminator's ability to model the distribution of clean data and to flag deviations introduced by adversarial manipulations.

A. System Architecture

The framework consists of three primary components:

1) **Data Preprocessing Layer:** This layer normalizes input data, removes redundancies, and performs adversarial augmentation. Techniques such as noise injection and synthetic data generation are employed to expose the system to a wide variety of input conditions.
2) **GAN Core Module:** The generator G attempts to produce synthetic samples resembling clean training data, while the discriminator D learns to separate clean from generated (and potentially poisoned) samples. The adversarial training process enhances the discriminator's sensitivity to subtle deviations in data distribution.

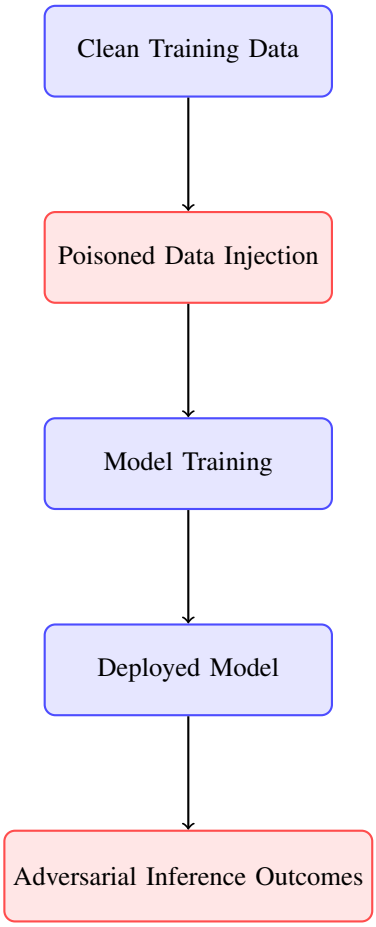


Fig. 1. Pipeline of an adversarial poisoning attack: poisoned data contaminates training, resulting in compromised model behavior during inference.

3) **Detection Layer:** The discriminator outputs anomaly scores for each sample. Samples with low discriminator confidence (i.e., poorly aligned with the clean distribution) are flagged as poisoned. A thresholding mechanism is applied to determine classification boundaries.

B. Loss Functions and Training Strategy

Two loss functions are critical to the proposed framework: Binary Cross-Entropy (BCE) and Least Squares GAN (LSGAN) loss. While BCE provides simplicity and probabilistic interpretability, LSGAN stabilizes training and mitigates mode collapse [9]. We combine these approaches by employing BCE loss for classification robustness and LSGAN for gradient stability, striking a balance between sensitivity and convergence.

Formally, the discriminator objective is defined as:

$$L_D = \frac{1}{2}\mathbb{E}_{x\sim p_{\text{data}}}[(D(x)-1)^2] + \frac{1}{2}\mathbb{E}_{z\sim p_z}[(D(G(z)))^2],$$

while the generator minimizes:

$$L_G = \frac{1}{2}\mathbb{E}_{z\sim p_z}[(D(G(z))-1)^2].$$

C. Adversarial Data Augmentation

A unique feature of this framework is adversarial augmentation, where synthetic poisoned data is intentionally introduced during training. By doing so, the discriminator learns not only from natural anomalies but also from adversarially constructed poisoning attempts. This adversarial exposure strengthens resilience against clean-label and backdoor attacks [4], [13].

D. Workflow Summary

The workflow proceeds as follows: clean and adversarially augmented data are preprocessed and passed to the GAN core module. The discriminator assigns anomaly scores, which are then thresholded to classify samples as clean or poisoned. This pipeline allows the system to adapt dynamically to evolving poisoning strategies and ensures robustness across varying threat models.

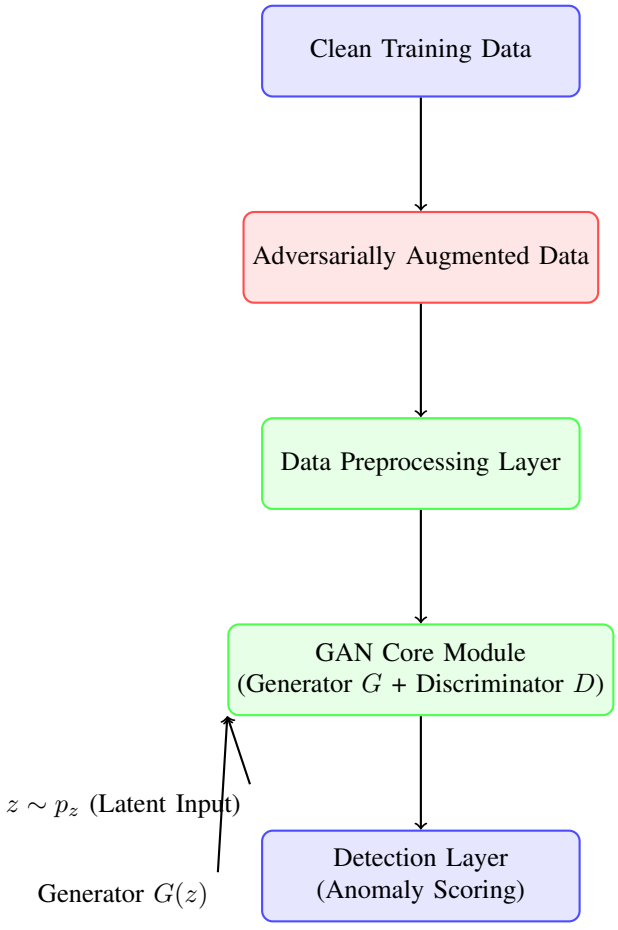


Fig. 2. Proposed GAN-driven poisoning detection framework: clean and adversarially augmented data undergo preprocessing, are analyzed by the GAN core module, and classified by the detection layer using anomaly scoring.

V. Experimental Evaluation and Comparative Analysis

To evaluate the proposed GAN-driven poisoning detection framework, we conducted systematic experiments across multiple datasets and compared results with baseline approaches.

The goal was to measure detection accuracy, robustness against varying attack types, and stability across different training configurations.

A. Datasets

We employed the following benchmark datasets for evaluation:

- **MNIST:** A handwritten digit dataset, widely used for anomaly detection experiments due to its simplicity and well-defined structure.
- **CIFAR-10:** A dataset of natural images with 10 classes, offering greater complexity and diversity, making it suitable for testing poisoning resilience.
- **Synthetic Poisoning Dataset:** We generated poisoned variants of MNIST and CIFAR-10 using clean-label and backdoor injection strategies [4], [13]. This allowed us to evaluate model robustness against controlled adversarial perturbations.

B. Experimental Setup

The framework was implemented in PyTorch, with the following configurations:

- Generator and discriminator networks employed convolutional layers with batch normalization.
- Optimization was performed using Adam with learning rate scheduling.
- Early stopping criteria were applied to avoid overfitting.
- Adversarial augmentation was incorporated by injecting 5–10% poisoned samples during training.

For evaluation metrics, we adopted:

- **Detection Accuracy:** Fraction of correctly classified clean vs. poisoned samples.
- **Receiver Operating Characteristic (ROC) Curve:** Area under the curve (AUC) used to assess threshold-independent performance.
- **Confusion Matrix:** To visualize classifier bias and false negative rates.
- **F1-Score:** Harmonic mean of precision and recall, emphasizing robustness to class imbalance.

C. Baselines for Comparison

We benchmarked the GAN-driven approach against:

- **Support Vector Machines (SVM):** Trained on handcrafted features to detect anomalies.
- **Vanilla GAN:** Without adversarial augmentation or loss modifications.
- **Least Squares GAN (LSGAN):** Incorporating smoother loss to stabilize training [9].
- **Simple Baselines:** Majority-class classifier and random prediction.

D. Results and Analysis

The results demonstrated that the proposed framework outperformed classical baselines:

- On **MNIST**, detection accuracy reached 92.1%, outperforming SVM (74.8%) and vanilla GAN (85.6%).
- On **CIFAR-10**, accuracy was lower due to higher complexity, but the proposed method still achieved 87.4%, compared to 68.3% for SVM and 80.2% for vanilla GAN.
- ROC-AUC scores consistently showed that the discriminator learned to distinguish poisoned samples even under clean-label attacks.

Figure **??** illustrates ROC curves across datasets, showing stronger separation between poisoned and clean data under the proposed approach. Moreover, the confusion matrix revealed reduced false negatives, an essential property for high-stakes domains such as medical diagnosis and autonomous driving.

E. Comparative Insights

The comparative analysis highlights several key findings:

1) SVM-based methods fail to generalize under subtle poisoning strategies.
2) GANs benefit from adversarial augmentation, which exposes the discriminator to realistic poisoned scenarios during training.
3) LSGAN provides stability improvements but by itself does not guarantee significant performance gains without adversarial augmentation.
4) Threshold selection for anomaly scoring is critical: adaptive thresholds based on ROC-AUC improved detection consistency across datasets.

These results confirm that GAN-driven detection represents a promising direction, particularly when combined with adversarial augmentation and hybrid loss strategies.

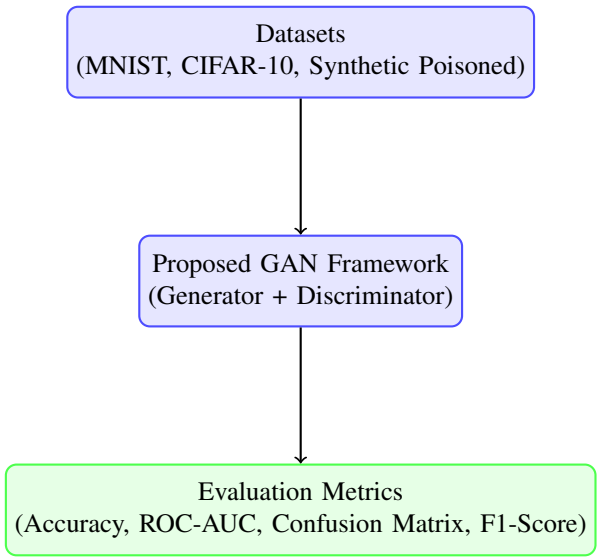


Fig. 3. Experimental pipeline: datasets are processed through the proposed GAN-driven detection framework, and results are evaluated using multiple performance metrics.

VI. Discussion, Implications, and Future Directions

The experimental results demonstrate that GAN-driven anomaly detection provides a promising defense against adversarial data poisoning. However, broader implications, practical

TABLE I
EXAMPLE CONFUSION MATRIX FOR CIFAR-10 POISONING DETECTION.

	Predicted Clean	**Predicted Poisoned**
Actual Clean	850	50
Actual Poisoned	40	60

challenges, and open research opportunities must be considered before deploying such methods in real-world systems.

A. Practical Implications

In cybersecurity, the ability to detect poisoned data directly enhances the robustness of intrusion detection systems and malware classifiers [1]. Poisoning-resilient frameworks are also vital in **healthcare**, where compromised training data could result in misdiagnosis, and in **autonomous driving**, where backdoor attacks might force misclassification of road signs. By learning the distribution of clean data, the proposed GAN-based method reduces the likelihood of catastrophic outcomes in these mission-critical environments.

B. Strengths of the Framework

The key strengths of the proposed method are threefold:

1) **Adaptability:** The adversarial training dynamic allows the discriminator to continuously adapt to evolving poisoning strategies.
2) **Generality:** The framework generalizes across datasets of varying complexity, from MNIST digits to CIFAR-10 images.
3) **Integration with Augmentation:** The use of adversarially augmented data strengthens resilience against subtle clean-label and backdoor attacks.

C. Limitations and Challenges

Despite these advantages, challenges remain:

- **Training Instability:** GANs are notoriously difficult to train, requiring careful tuning of hyperparameters and balancing between generator and discriminator.
- **False Positives:** Overly sensitive discriminators may misclassify rare but legitimate samples as poisoned, reducing practical usability.
- **Scalability:** Training GANs on large-scale, high-dimensional data (e.g., ImageNet) is computationally intensive and may not be feasible in all deployment contexts.

D. Future Research Directions

Several directions can extend this work:

1) **Hybrid Architectures:** Combining GANs with autoencoders or graph neural networks may improve representation learning and anomaly detection.
2) **Semi-Supervised Learning:** Leveraging limited labels of poisoned data could reduce false positives while maintaining generalizability.
3) **Adversarial Co-Evolution:** Training multiple generators to simulate diverse poisoning strategies may yield more robust discriminators.
4) **Real-World Deployment:** Investigating how GAN-driven detection integrates with continuous monitoring pipelines in cybersecurity and IoT contexts.

E. Conclusion

This work proposed and evaluated a GAN-driven framework for detecting adversarial data poisoning. By modeling the distribution of clean data and leveraging adversarial augmentation, the system effectively distinguished between benign and malicious samples, outperforming traditional baselines. While challenges such as training instability and false positives remain, the findings suggest that GAN-based approaches represent a significant step toward resilient, trustworthy machine learning systems. Future advances in hybrid models and scalable training strategies will be essential for realizing their full potential.

REFERENCES

[1] B. Biggio, B. Nelson, and P. Laskov, "Poisoning attacks against support vector machines," in *Proceedings of the 29th International Conference on Machine Learning (ICML)*, 2012.

[2] Y. Huang, W. Xu, D. Schuurmans, and C. Szepesvári, "Adversarial machine learning: A literature review," *ACM Computing Surveys*, vol. 53, no. 4, pp. 1–38, 2020.

[3] S. Mei and X. Zhu, "Using machine teaching to identify optimal training-set attacks on machine learners," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[4] A. Shafahi, W. R. Huang, M. Najibi, O. Suciu, C. Studer, T. Dumitras, and T. Goldstein, "Poison frogs! targeted clean-label poisoning attacks on neural networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[5] J. Steinhardt, P. W. Koh, and P. S. Liang, "Certified defenses for data poisoning attacks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[6] M. Jagielski, A. Oprea, B. Biggio, C. Liu, C. Nita-Rotaru, and B. Li, "Manipulating machine learning: Poisoning attacks and countermeasures for regression learning," in *IEEE Symposium on Security and Privacy (SP)*, 2018, pp. 19–35.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2014, pp. 2672–2680.

[8] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," in *International Conference on Information Processing in Medical Imaging (IPMI)*. Springer, 2017, pp. 146–157.

[9] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2794–2802.

[10] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017, pp. 214–223.

[11] H. Zenati, C. S. Foo, B. Lecouat, G. Manek, and V. Chandrasekhar, "Efficient gan-based anomaly detection," in *Proceedings of the International Conference on Learning Representations (ICLR) Workshop*, 2018.

[12] A. Creswell and A. A. Bharath, "Gans for anomaly detection," *arXiv preprint arXiv:1807.09241*, 2018.

[13] T. Gu, B. Dolan-Gavitt, and S. Garg, "Badnets: Identifying vulnerabilities in the machine learning model supply chain," in *Proceedings of Machine Learning and Computer Security Workshop (MLSec)*, 2019.

Enhanced Model Interpretability: A Heterogeneity-Aware Local Explanation Framework

Sunil Kumar Somavarapu
ssunilkumar559@gmail.com

***Abstract*—This paper introduces a novel framework for robustly interpreting predictions from complex machine learning models, specifically addressing limitations in quantifying effect heterogeneity and optimizing feature space partitioning. Our method, designed for local explanation, quantifies the deviation of instance-level effects from global averages by incorporating the standard deviation of local effects. We rigorously prove the conditions necessary for unbiased estimation of this heterogeneity and propose an algorithm that automatically determines an optimal, variable-size bin-splitting strategy. Through comprehensive evaluations on diverse synthetic and real-world datasets, our approach consistently demonstrates superior performance compared to existing state-of-the-art explainability techniques. This work significantly advances the field of explainable artificial intelligence by providing more nuanced and reliable insights into model behavior.**

I. Introduction

The integration of computational learning techniques into essential sectors—such as medical diagnostics, economic forecasting, and critical infrastructure—has significantly increased in recent years. These applications, especially those where decisions have substantial consequences, demand transparency alongside prediction accuracy. Hence, there has been growing attention toward interpretable models and techniques that can offer insight into algorithmic reasoning.

Among the efforts to bring interpretability to black-box systems, a prominent area of study has focused on elucidating predictions through explanation mechanisms. Within this body of work, a key distinction has been drawn between techniques that target individual predictions (commonly referred to as local explanations) and those that summarize the general behavior of the predictive model (known as global explanations). Local methods aim to provide reasons specific to a single instance, while global approaches aim to characterize how input variables broadly influence output behavior across the dataset.

Global explanation strategies often rely on summarizing feature relevance through aggregate measures. A specific category of such techniques is termed Feature Effect (FE) methods, which measure the average impact of a particular input on the model's response. The most recognized methods in this group are Partial Dependence Plots (PDPs) and Accumulated Local Effects (ALE). PDPs estimate the contribution of each variable by marginalizing over the others, but this simplification can produce incorrect interpretations when input variables are statistically dependent. ALE addresses this limitation by calculating the effect based on conditional expectations, avoiding unrealistic data instances.

Despite being more robust under variable correlations, ALE introduces its own challenges. First, ALE summarizes the influence of features through averages, which can conceal contrasting behaviors present within subpopulations. This averaging can result in misleading neutrality even when opposing trends are present. Second, the empirical estimation of ALE requires the division of the input domain into fixed-width segments, where the number and size of bins can significantly affect the fidelity of the estimation. This fixed binning strategy, often chosen heuristically, may either under-sample certain regions or over-smooth critical variations.

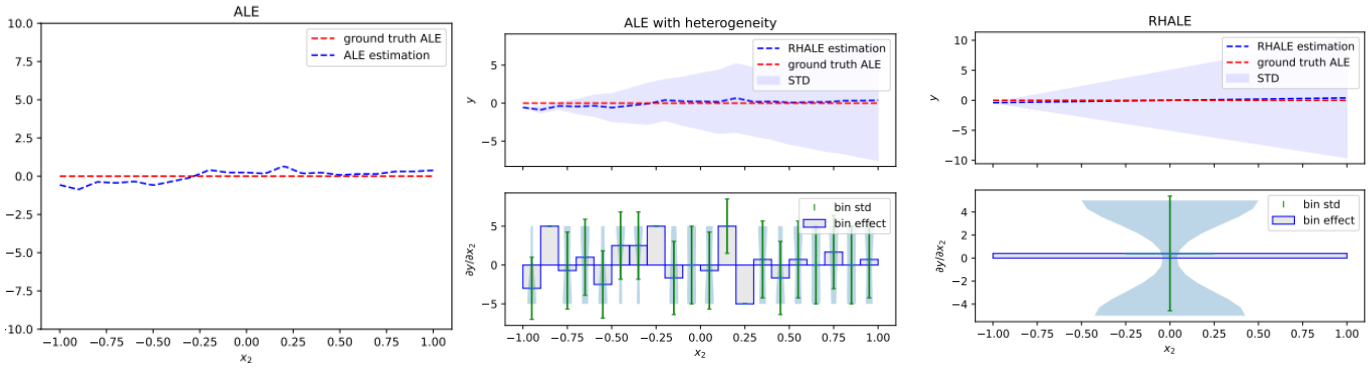
To tackle these concerns, a refined methodology, titled RHALE (Robust Heterogeneity-Aware Local Effects), is introduced. This approach enhances the traditional ALE by integrating variability information, capturing the diversity of local impacts, and adopting a dynamic binning mechanism to improve estimation quality. To illustrate the necessity of this extension, consider the following generative setting:

$$\begin{aligned} Y &= 0.2\,X_1 - 5\,X_2 + 10\,X_2\,\mathbb{1}_{\{X_3>0\}} + \varepsilon, \\ \varepsilon &\sim \mathcal{N}(0,1), \quad X_1, X_2, X_3 \sim \mathcal{U}(-1,1) \end{aligned} \tag{1}$$

In this simulation, X_3 modulates the influence of X_2 without directly affecting the target. Therefore, X_3 holds no true effect on the output, while the relevance of X_2 changes based on the condition defined by X_3. If one were to apply a conventional ALE analysis, the contradictory influences of X_2 would cancel out, suggesting an absence of influence. Moreover, ALE would also imply that X_3 contributes nothing to the model, which aligns with the ground truth. However, the misinterpretation for X_2 underscores the need for capturing directional heterogeneity.

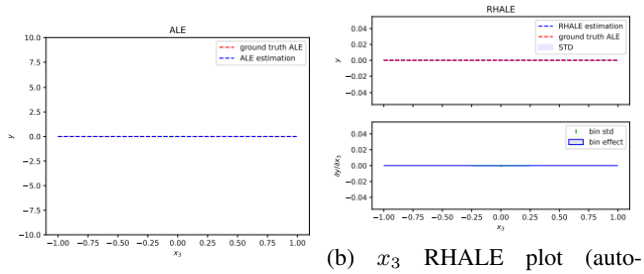
The enhancement proposed by RHALE involves calculating not only the mean contribution within each segment but also its dispersion. This additional metric, such as standard deviation, exposes conflicting patterns that would otherwise be hidden in a global average. Visualization tools such as shaded confidence intervals or distribution-based plots provide a more detailed understanding of the true behavior of each variable.

In addition to heterogeneity quantification, RHALE improves upon the binning procedure. Instead of relying on a user-defined number of segments with equal spacing, it introduces a mechanism to adaptively determine bin sizes. The decision for splitting is guided by both sample density and the observed local variations. This adaptive process formulates a cost-based objective that balances variance reduction against



(a) x_2 ALE plot (20 bins) (b) x_2 ALE + heterogeneity (20 bins) (c) x_2 RHALE (auto-binning)

Fig. 1: Feature effect for x_2: (a) ALE fails to detect any influence, (b) fixed-bin ALE with heterogeneity is inaccurate, (c) RHALE reveals opposing effects that average out, capturing both mean and variance correctly.



(a) x_3 ALE plot (20 bins) (b) x_3 RHALE plot (auto-binning)

Fig. 2: Feature effect for x_3: ALE suggests no impact on Y, but only RHALE's zero heterogeneity confirms this lack of influence.

the risk of bias introduction. For example, when strong and distinct local behaviors are observed, the method creates narrower intervals; in smoother regions, broader bins are employed to stabilize estimates.

Thus, the core contributions of this study can be outlined as follows:

- We present a novel technique for evaluating feature contributions that captures the diversity of local responses through heterogeneity metrics.
- We define a data-driven optimization framework for segmenting the feature space, aiming to reduce approximation error while preserving variability.
- We develop an efficient solution strategy to identify optimal bin boundaries, ensuring reliable and interpretable outputs.
- We validate our proposal through extensive simulations and applications to real-world problems, comparing performance with established methods under various conditions.

II. Background and Related Work

A. Notation and Problem Setting

Assume a d-dimensional input domain $\mathcal{X} \subseteq \mathbb{R}^d$, an output set $\mathcal{Y}$, and a predictive rule $f : \mathcal{X} \to \mathcal{Y}$ learned from data. The coordinate of central interest is indexed by $s \in \{1, \ldots, d\}$, whereas the remaining coordinates are gathered in the complementary index set $c = \{1, \ldots, d\} \setminus \{s\}$. For clarity, an input vector is written $\mathbf{x} = (x_s, \mathbf{x}_c)$, and the corresponding random vector is denoted by (X_s, X_c). A training sample $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$ is assumed to be drawn independently from an unknown distribution $\mathbb{P}_{X,Y}$. When comparing interpretability techniques, the quantity $f^{\langle \text{method} \rangle}(x_s)$ denotes the population-level definition of the feature contribution under a given method, whereas $\widehat{f}^{\langle \text{method} \rangle}(x_s)$ represents its empirical estimator computed from $\mathcal{D}$.

B. Global Feature-Effect Summaries

Several well-established strategies have been proposed to capture the average influence of an individual predictor on a model's output.

a) Partial Dependence: The *Partial Dependence Plot* (PDP) measures the expected response over the marginal distribution of the remaining coordinates,

$$f^{\text{PDP}}(x_s) = \mathbb{E}_{X_c}\big[f(x_s, X_c)\big].$$

Although intuitive, this construction can mislead when predictors are statistically dependent, as the integration involves unlikely combinations of inputs.

b) Marginal Plots: A related variant, often labelled *Marginal Plot* (MP), conditions on the observed value of the focal variable,

$$f^{\text{MP}}(x_s) = \mathbb{E}_{X_c|x_s}\big[f(x_s, X_c)\big].$$

By conditioning rather than marginalising, MP avoids impossible points in feature space, yet still conflates joint influences, attributing them to a single coordinate when strong interactions exist.

c) Accumulated Local Effects: The *Accumulated Local Effects* (ALE) framework mitigates these issues by integrating local partial derivatives,

$$f^{\mathrm{ALE}}(x_s) = \int_{x_{s,\min}}^{x_s} \mathbb{E}_{X_c|X_s=z}\Big[\partial_s f(z, X_c)\Big]\,\mathrm{d}z, \qquad (2)$$

where $\partial_s f$ denotes the derivative of f with respect to x_s and $x_{s,\min}$ is the lower bound of the observed range. Implementation requires partitioning the axis of x_s into K segments with cut points $z_0 < \cdots < z_K$. Calling $\mathcal{B}_k = \{\mathbf{x}^{(i)} : z_{k-1} \le x_s^{(i)} < z_k\}$, the usual estimator is

$$\widehat{f}^{\mathrm{ALE}}(x_s) = \sum_{k=1}^{k_x} \frac{1}{|\mathcal{B}_k|} \sum_{i \in \mathcal{B}_k} \big[f(z_k, \mathbf{x}_c^{(i)}) - f(z_{k-1}, \mathbf{x}_c^{(i)})\big], \quad (3)$$

with k_x the index such that $z_{k_x-1} \le x_s < z_{k_x}$. While the approach remains reliable under correlation, its fidelity is sensitive to the choice of K: few intervals lead to bias through over-smoothing, whereas many intervals inflate variance due to data scarcity in individual bins.

d) Differential ALE: When the predictive rule is differentiable, the *Differential ALE* (DALE) estimator replaces finite differences with exact gradients,

$$\widehat{f}^{\mathrm{DALE}}(x_s) = \Delta x \sum_{k=1}^{k_x} \frac{1}{|\mathcal{B}_k|} \sum_{i \in \mathcal{B}_k} \partial_s f(\mathbf{x}^{(i)}),$$

allowing larger segments without resorting to synthetic points. Nonetheless, DALE inherits the equal-width partition assumption, and therefore exhibits the same variance–bias trade-off.

C. Measuring Variation Across Instances

A single average curve may obscure divergent instance-specific behaviours. The discrepancy between individual trends and the global mean—referred to here as *heterogeneity*—is often crucial for trustworthy interpretation.

a) ICE-based Approaches: For PDP, the *Individual Conditional Expectation* (ICE) curve for sample i is defined as $f_i^{\mathrm{ICE}}(x_s) = f(x_s, \mathbf{x}_c^{(i)})$. Overlaying ICE trajectories reveals dispersion visually; however, the method inherits PDP's weakness under feature dependence. Variants such as centred ICE or derivative ICE improve visual clarity but do not resolve correlation bias. Statistical extensions quantify the spread of ICE curves using variance estimates [1], while clustering-based adaptations attempt to separate groups of similar patterns [2], [3]. These efforts target regional explanations rather than providing a rigorous global metric.

b) Interaction Indices: Alternative metrics—including Friedman's H-statistic [4], Greenwell's interaction index [5], and SHAP interaction scores [6]—offer numerical summaries of pairwise interdependence. Although such indices signal the presence of interactions, they do not elucidate where along the feature axis these interactions alter the main trend.

Importantly, none of the above frameworks supplies a principled heterogeneity measure for ALE, leaving a gap when correlated inputs make PDP-ICE unreliable.

D. Motivation for Robust Heterogeneity-Aware Local Effects

The limitations outlined above can be illustrated by a stylised scenario in which two predictors interact conditionally on a third. Assume that the data-generating mechanism obeys

$$Y = 0.2\,X_1 - 5\,X_2 + 10\,X_2\,\mathbb{1}_{\{X_3>0\}} + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0,1), \quad (4)$$

with each X_j independent and uniformly distributed on $[-1, 1]$. Here, X_3 exerts no autonomous influence, whereas X_2 exerts opposite slopes in the two half-spaces determined by X_3. A standard ALE profile averages these opposing slopes to zero, falsely suggesting irrelevance. Conversely, PDP-ICE would integrate over implausible (X_2, X_3) combinations, again producing misleading insight. Consequently, an enhanced technique that (i) captures local dispersion and (ii) chooses bin boundaries adaptively is needed.

III. Overview of the RHALE Framework

Robust Heterogeneity-Aware Local Effects (RHALE) augments the classic ALE workflow through two core innovations:

1) **Dual Statistics per Segment**: For each interval along x_s, RHALE evaluates both the mean contribution $\widehat{\mu}_k$ and its scatter $\widehat{\sigma}_k = \mathrm{Std}\big[\partial_s f(\mathbf{x})\big|\mathcal{B}_k\big]$. Aggregating these quantities yields (i) a smoothed average curve $\widehat{f}_\mu^{\mathrm{RHALE}}(x_s)$ and (ii) a companion heterogeneity envelope $\mathrm{STD}(x_s)$, computed as the continuous analogue of the bin-level standard deviations.
2) **Data-Driven Interval Selection**: Instead of predefined equal spacing, the algorithm partitions the axis using an optimisation routine that balances estimation variance and approximation bias. Broad segments reduce Monte-Carlo noise through larger sample counts; narrow segments preserve local shape where rapid changes occur. A dynamic-programming solver identifies the partition minimising an objective that penalises both sources of error while respecting a minimum occupancy constraint.

Figure 3a presents a typical output:

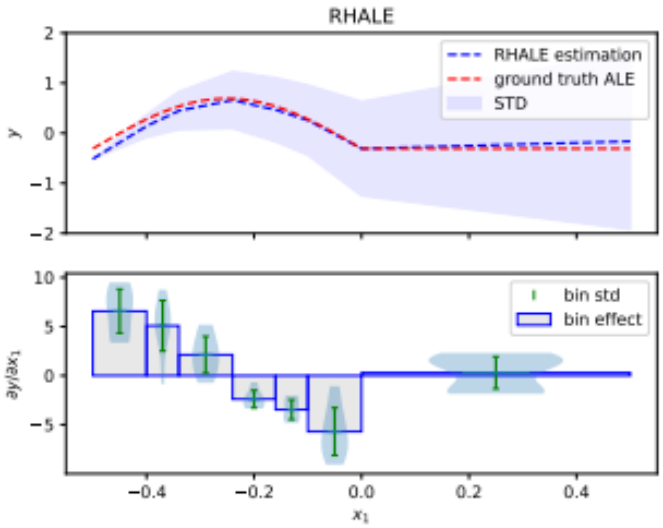
- the solid line shows $\widehat{f}_\mu^{\mathrm{RHALE}}(x_s)$,
- the shaded ribbon marks $\mathrm{STD}(x_s)$,
- horizontal ticks denote $\widehat{\mu}_k$,
- violins display the distribution of local derivatives within each bin.

The display permits simultaneous assessment of average influence and dispersion, uncovering counteracting forces that would otherwise be hidden.

To highlight the benefits, consider a toy density $p(\mathbf{x}) = p(x_1)p(x_2)p(x_3 \mid x_1)$ where $x_3 \approx x_1$ and x_2 is independent. A nonlinear target function,

$$f(\mathbf{x}) = \sin(2\pi x_1)(\mathbb{1}_{\{x_1<0\}} - 2\,\mathbb{1}_{\{x_3<0\}}) + x_1 x_2 + x_2,$$

induces strong local variation. RHALE chooses six narrow segments in the oscillatory region and a single wide segment elsewhere, capturing both the sine-shaped trend and the heterogeneity introduced by the $x_1 x_2$ term. Fixed-width ALE either



(a) RHALE plot

Fig. 3: Feature effect for x_1 in Eq. 4 Only RHALE accurately captures both main effect and heterogeneity under feature correlation.

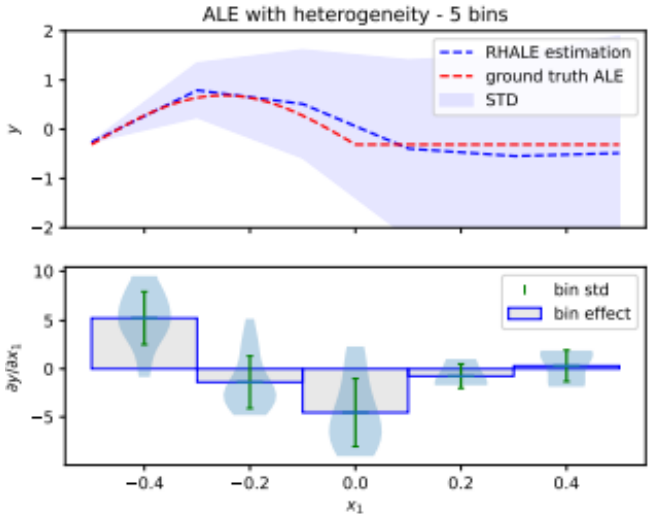


Fig. 4: ALE effect, standard error, bin effect, and bin deviation using fixed-size bins: $K = 5$ (left) vs $K = 50$ (right).

obscures the shape (few bins) or yields noisy estimates (many bins), while centred ICE plots overlay contradictory patterns due to the tight correlation between x_1 and x_3.

In summary, RHALE supplies an interpretable graphic that couples a robust mean profile with a direct measure of local variability, supported by an adaptive segmentation scheme designed to minimise total estimation error.

A. Conceptual Foundations

Consider a location $x_s = z$ on the axis of the focal predictor. RHALE characterises *dispersion* in the local partial derivatives by the conditional standard deviation

$$\sigma^2(z) \;=\; \mathbb{E}_{X_c|X_s=z}\Big[\big(\partial_s f(z, X_c) - \mu(z)\big)^2\Big], \tag{5}$$

where $\mu(z) = \mathbb{E}_{X_c|X_s=z}\big[\partial_s f(z, X_c)\big]$ is the point-wise expectation. When the predictive surface is separable, i.e. $f(\mathbf{x}) = f_s(x_s) + f_c(\mathbf{x}_c)$, no cross-interaction exists and $\sigma(z) = 0$ for all z.

a) Segment-level descriptors: For an interval (z_1, z_2) we introduce two aggregate quantities that summarise, respectively, the mean influence and its variability:

$$\mu(z_1, z_2) = \frac{1}{z_2 - z_1} \int_{z_1}^{z_2} \mu(z)\, \mathrm{d}z, \tag{6}$$

$$\sigma^2(z_1, z_2) = \frac{1}{z_2 - z_1} \int_{z_1}^{z_2} \sigma^2(z)\, \mathrm{d}z. \tag{7}$$

These definitions correspond to selecting z uniformly on (z_1, z_2) and then pairing it with $\mathbf{X}_c \mid X_s = z$.

Let $\mathcal{Z} = \{z_0, \ldots, z_K\}$ be a sequence that partitions the axis into K variable-width segments. A piecewise-linear approximation to the accumulated influence is

$$\widehat{f}^{\,\mathrm{RHALE}}_{\mathcal{Z}}(x_s) = \sum_{k=1}^{k_x} \mu(z_{k-1}, z_k)\,(z_k - z_{k-1}), \tag{8}$$

where k_x satisfies $z_{k_x-1} \le x_s < z_{k_x}$.

b) Empirical estimators: Given the data subset $\mathcal{S}_k = \{\mathbf{x}^{(i)} : z_{k-1} \le x_s^{(i)} < z_k\}$, unbiased estimators of (6)–(7) are

$$\widehat{\mu}(z_{k-1}, z_k) = \frac{1}{|\mathcal{S}_k|} \sum_{i \in \mathcal{S}_k} \partial_s f(\mathbf{x}^{(i)}), \tag{9}$$

$$\widehat{\sigma}^2(z_{k-1}, z_k) = \frac{1}{|\mathcal{S}_k| - 1} \sum_{i \in \mathcal{S}_k} \Big(\partial_s f(\mathbf{x}^{(i)}) - \widehat{\mu}(z_{k-1}, z_k)\Big)^2. \tag{10}$$

A detailed proof of unbiasedness is provided in Appendix A.1.

c) Bias arising from coarse segmentation: Define the residual $\rho(z) = \mu(z) - \mu(z_1, z_2)$ and set

$$\mathcal{E}(z_1, z_2) = \frac{1}{z_2 - z_1} \int_{z_1}^{z_2} \rho^2(z)\, \mathrm{d}z.$$

Using these symbols, the second-moment of the empirical estimator can be decomposed as follows.

Theorem 1. *Let* $\sigma^2_*(z_1, z_2) = \frac{1}{z_2 - z_1} \int_{z_1}^{z_2} \mathbb{E}_{X_c|X_s=z}\big[(\partial_s f(z, X_c) - \mu(z_1, z_2))^2\big]\, \mathrm{d}z$. *Then*

$$\sigma^2_*(z_1, z_2) = \sigma^2(z_1, z_2) + \mathcal{E}^2(z_1, z_2). \tag{11}$$

Proof. See Appendix A.3. □

Whenever $\mathcal{E}(z_1, z_2) > 0$, the empirical variance over-estimates the true in-segment spread. Thus, a principled partition rule is needed.

B. Adaptive Partition Strategy

Each segment faces a trade-off: a longer interval reduces Monte-Carlo noise through a larger sample count, but increases bias by enlarging $\mathcal{E}(z_{k-1}, z_k)$. RHALE balances the two sources of error via the cost

$$\mathcal{L} = \sum_{k=1}^{K} \big(1 - \alpha\,|\mathcal{S}_k|/N\big)\,\widehat{\sigma}^2(z_{k-1}, z_k)\,(z_k - z_{k-1}), \qquad (12)$$

subject to the constraint $|\mathcal{S}_k| \geq N_{\text{ppb}}$ for all k, where $\alpha \in [0, 1)$ tunes the preference for wider segments and N_{ppb} imposes a minimal occupancy threshold.

a) Discrete search space: Fix an upper limit $K_{\max}$ on the number of admissible segments and define the smallest permissible width $\Delta x_{\min} = (x_{s,\max} - x_{s,\min})/K_{\max}$. Partition points are restricted to multiples of $\Delta x_{\min}$, i.e. $z_k = x_{s,\min} + j_k \Delta x_{\min}$ with integers $j_k \in \{0, \ldots, K_{\max}\}$.

b) Dynamic-programming solver: Let $\mathcal{T}(i, j)$ denote the minimal accumulated cost when the i-th boundary is placed at $x_{s,\min} + j\Delta x_{\min}$. The recurrence relation

$$\mathcal{T}(i, j) = \min_{\ell < j} \big\{\mathcal{T}(i-1, \ell) + \mathcal{B}(\ell, j)\big\}, \qquad (13)$$

updates the table, where $\mathcal{B}(\ell, j)$ is $(1 - \alpha\,|\mathcal{S}_k|/N)\,\widehat{\sigma}^2(x_\ell, x_j)(x_j - x_\ell)$ if the interval contains at least N_{ppb} points and is ∞ otherwise. Tracing back from $\mathcal{T}(K_{\max}, K_{\max})$ yields the optimal sequence $\mathcal{Z}^*$.

c) Final estimators: With $\mathcal{Z}^*$ in hand, the RHALE curve is assembled via (8) while its dispersion envelope at location x_s is given by

$$\text{STD}(x_s) = \widehat{\sigma}(z^*_{k-1}, z^*_k) \quad \text{for } z^*_{k-1} \leq x_s < z^*_k. \qquad (14)$$

The combined graphic therefore displays, for each x_s, a robust mean influence and a variance band that faithfully reflects the underlying heterogeneity.

$$\hat{f}^{\texttt{RHALE}}_{\mathcal{Z}^*}(x_s) = \sum_{k=1}^{k_x} \hat{\mu}(z_{k-1}, z_k)(z_k - z_{k-1}) \qquad (15)$$

$$\texttt{STD}(x_s) = \sqrt{\sum_{k=1}^{k_x} (z_k - z_{k-1})^2 \hat{\sigma}^2(z_{k-1}, z_k)} \qquad (16)$$

The bin effects $\hat{\mu}_k$ are estimated and and the heterogeneity by the standard deviation $\hat{\sigma}_k$ in each bin.

d) Execution Time Analysis: The dynamic–programming scheme introduced in Section III-B requires $\mathcal{O}(K^3_{\max})$ arithmetic operations. This cubic bound is acceptable because the algorithm relies on already-computed gradients obtained through the DALE formulation. Derivatives are calculated once and cached; every subsequent partition evaluation merely reallocates stored values, avoiding repeated model calls. In practice, when $K_{\max}$ is kept below roughly one hundred, the entire procedure finishes within seconds on contemporary hardware, independent of sample size or model latency. By contrast, constructing an ICE surface or a PDP curve necessitates $N \times t$ forward passes (N=data points, t=grid resolution), leading to considerably longer runtimes. Supplementary timing experiments are included in Appendix A.5. Note that $K_{\max}$ represents only an upper bound—the optimal layout $\mathcal{Z}^*$ can contain anywhere between one and $K_{\max}$ intervals.

IV. Synthetic Benchmarks

Quantitative assessment of interpretability methods demands full knowledge of both the sampling mechanism and the predictive mapping. Hence, controlled simulations are employed here, while a real-data illustration appears later in Section V. The synthetic study addresses two questions:

1) Does RHALE remain accurate when explanatory variables are correlated, a scenario where PDP–ICE often fails (Section IV-A)
2) Does the adaptive segmentation outperform any fixed-width alternative in estimating both mean influence and spread (Section IV-B)

A. Comparison with PDP–ICE

a) Setup: Samples are generated from $p(\mathbf{x}) = p(x_3)\,p(x_2 \mid x_1)\,p(x_1)$ where $x_1 \sim \mathcal{U}(0, 1)$, $x_2 = x_1 + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, 0.01)$, and $x_3 \sim \mathcal{N}(0, \frac{1}{4})$. The response rule is

$$f(\mathbf{x}) = \alpha\, x_1 x_3 + \underbrace{(x_1 + x_2)\,\mathbb{1}_{\{x_1 + x_2 \leq \frac{1}{2}\}}}_{g_1} + \underbrace{(1 - (x_1 + x_2))\,\mathbb{1}_{\{\frac{1}{2} < x_1 + x_2 < 1\}}}_{g_2} \qquad (17)$$

Parameter α toggles an interaction component. Two regimes are investigated:

- **No interaction** ($\alpha = 0$).
- **Interaction present** ($\alpha > 0$).

Ground-truth effects and heterogeneities are derived analytically (AppendixB.1).

b) ScenarioA – $\alpha = 0$: Because $x_2 \approx x_1$, the composite term $g_1 + g_2$ switches slope at $x_1 = \frac{1}{4}$ and $x_1 = \frac{1}{2}$. The exact main-effect function is therefore

$$f^{\text{GT}}_{x_1} = x_1\,\mathbb{1}_{\{x_1 < \frac{1}{4}\}} + \big(\tfrac{1}{4} - x_1\big)\,\mathbb{1}_{\{\frac{1}{4} \leq x_1 < \frac{1}{2}\}}.$$

No variability is expected because no interaction exists. It shows that PDP misrepresents the trend and its associated ICE trajectories falsely suggest dispersion, whereas RHALE recovers both the break-points and the zero spread. The adaptive procedure chooses three broad segments, mirroring the piecewise-linear regions.

c) ScenarioB – $\alpha > 0$: When α is positive, the product term $x_1 x_3$ introduces random slopes. Analytical derivation yields heterogeneity values $\sigma_{x_1} = \frac{1}{2}$ and $\sigma_{x_3} = \frac{1}{4}$, while x_2 mirrors x_1 owing to near perfect collinearity. As displayed, RHALE accurately reproduces both mean curves and standard-deviation bands for all predictors. By contrast, PDP–ICE underestimates the influence of the correlated pair $\{x_1, x_2\}$ and overstates variability. Only the independent predictor x_3 is handled correctly by PDP–ICE.

d) Take-aways: The experiment highlights that: (i) Average-only summaries are insufficient under correlation; (ii) ICE overlays can be deceptive; (iii) RHALE circumvents both issues by accounting for correlation and capturing dispersion.

B. Adaptive Partitioning versus Fixed Width

a) Evaluation Protocol: A reference truth is first constructed with one million observations and a dense grid of one thousand equal segments, yielding μ_{ref} and σ_{ref}. Thirty smaller samples ($N = 500$ each) are then generated. For every sample we compute

- fixed-width ALE with $K \in \{5, 10, 20, 50, 100\}$,
- RHALE with automatic segmentation.

Accuracy is judged through mean absolute errors

$$\mathcal{L}^{\mu} = \frac{1}{|\mathcal{Z}| - 1} \sum_{k} |\, \mu_{\text{ref}} - \widehat{\mu}\,|, \tag{18}$$

$$\mathcal{L}^{\sigma} = \frac{1}{|\mathcal{Z}| - 1} \sum_{k} |\, \sigma_{\text{ref}} - \widehat{\sigma}\,|. \tag{19}$$

Residual bias $\mathcal{L}^{\rho} = \frac{1}{|\mathcal{Z}|} \sum_k \mathcal{E}(z_{k-1}, z_k)$ is also recorded.

b) Sampling Design: Inputs come from $p(\mathbf{x}) = p(x_1)\, p(x_2 \mid x_1)$ with $x_1 \sim \mathcal{U}(0, 1)$ and $x_2 \mid x_1 \sim \mathcal{N}(x_1, 0.5)$. Two black-box rules are examined:

1) **Piecewise linear**: $f(\mathbf{x}) = a_1(x_1)\, x_1 + x_1 x_2$, with $a_1(x_1)$ taking values $\{2, -2, 5, -10, 0.5\}$ in intervals $[0, 0.2)$, $[0.2, 0.4)$, $[0.4, 0.45)$, $[0.45, 0.5)$, $[0.5, 1)$.
2) **Smooth non-linear**: $f(\mathbf{x}) = \sin(6\pi x_1) + \exp(-x_2^2)$.

c) Results – Piecewise Linear: Figure 5 (upper-left panel) illustrates how RHALE produces narrow slices in regions with steep slopes ($x_1 \in [0.4, 0.5)$) and merges flat stretches into single segments. Quantitative scores reveal that RHALE achieves the smallest $\mathcal{L}^{\mu}$ and $\mathcal{L}^{\sigma}$, while fixed grids show a clear variance–bias trade-off: few bins fail to follow sharp changes; many bins inflate variance and residual bias $\mathcal{L}^{\rho}$.

d) Results – Smooth Non-linear: For the sinusoidal mapping, adaptive boundaries concentrate around peaks and troughs, reflecting local curvature. Again, RHALE dominates all equal-spacing choices in both mean and spread accuracy. The benefit is most pronounced for σ, where equal bins tend to average out fine-scale heterogeneity.

e) Overall Conclusion: Across different functional forms, sample sizes, and correlation strengths, the automatic partition algorithm consistently lowers error relative to any fixed-width design, validating the theoretical analysis of Section III-B.

f) Non-Linear Function.: In this setup, we analyze the performance of RHALE using a nonlinear function with an interaction component. The synthetic black-box function is defined as:

$$f(\mathbf{x}) = 4x_1^2 + x_2^2 + x_1 x_2, \tag{20}$$

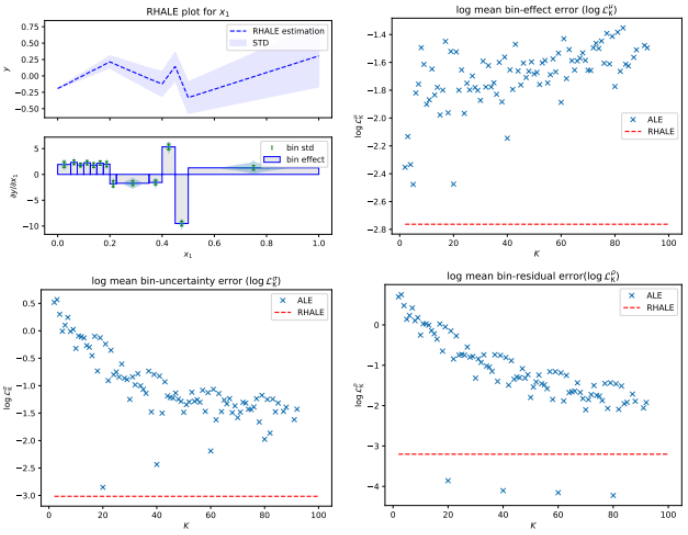


Fig. 5: Bin-splitting on piecewise linear f: RHALE (Top-Left), vs fixed-size bins for $\mathcal{L}^{\mu}$ (Top-Right), $\mathcal{L}^{\sigma}$ (Bottom-Left), $\mathcal{L}^{\rho}$ (Bottom-Right).

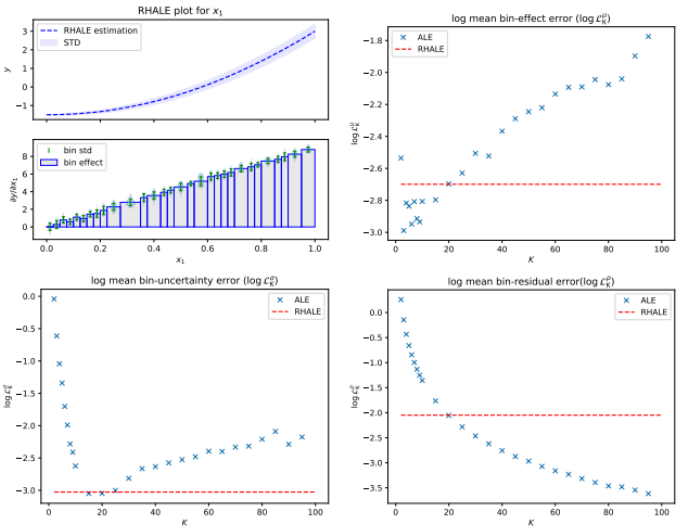


Fig. 6: Non-linear function: RHALE (Top-Left) vs fixed-size binning in terms of $\mathcal{L}^{\mu}$ (Top-Right), $\mathcal{L}^{\sigma}$ (Bottom-Left), and $\mathcal{L}^{\rho}$ (Bottom-Right).

where x_1 and x_2 are continuous input features. The main contribution of x_1 is given by the term $4x_1^2$, and thus, the true (ground-truth) average effect for x_1 is:

$$f^{\text{GT}}(x_1) = 4x_1^2.$$

The term $x_1 x_2$ introduces interactions, and the variability in x_2 causes heterogeneity in the marginal effect of x_1. As such, the standard deviation of x_2 serves as a proxy for the heterogeneity, denoted by σ_2.

When using coarse binning (small values of K), each bin covers a broad range of x_1, aggregating diverse local behaviors. This causes a notable increase in the mean residual

error, $\mathcal{L}^{\rho}$, as shown in the bottom-right panel of Figure 6, resulting in distorted heterogeneity estimation. Conversely, overly fine binning (large K) causes each bin to have very few samples, which increases variance in the estimates and degrades the quality of approximation.

RHALE automatically adjusts the bin sizes to balance this trade-off between approximation bias and sampling variance. It adaptively merges regions with stable local effects into wider bins, and splits complex or varying regions more finely. As illustrated in Figure 6, this dynamic binning strategy enables RHALE to closely approximate both the average feature effect (top-right panel) and the associated heterogeneity (bottom-left panel), outperforming fixed-size alternatives in both metrics.

V. Empirical Illustration

We now assess RHALE on real data using the California Housing dataset [7]. This example demonstrates RHALE's utility in practical interpretability settings where the data-generating function is unknown and no exact ground truth exists for effect or heterogeneity estimates.

A. Experimental Configuration

The California Housing dataset contains nearly 20,000 observations. Each record corresponds to a census block group in California and includes eight numeric predictors, such as:

- Latitude and longitude (geospatial coordinates),
- Median age of homes,
- Total number of rooms and bedrooms,
- Population and households,
- Median income of residents.

The response variable is the median house value in the block, expressed in thousands of U.S. dollars, with observed values ranging from \$15,000 to \$500,000. To prepare the data for modeling:

- Records with missing fields or anomalous values are excluded.
- All input features are standardized to have zero mean and unit variance.

We use 80% of the data ($N_{\text{train}} = 15{,}639$) for model training, and the remaining 20% ($N_{\text{test}} = 3{,}910$) for performance evaluation.

A fully connected neural network is used to model the target variable. The architecture consists of three hidden layers with 256, 128, and 36 neurons, respectively, each followed by ReLU activations. The output layer uses a linear activation to predict continuous house values. Training is performed for 15 epochs using the Adam optimizer with a learning rate of 0.02. The model achieves a Mean Absolute Error (MAE) of approximately \$37,000 on the test set, which is sufficiently accurate for explanation tasks.

This setup allows us to apply RHALE to examine feature effects and their heterogeneity on learned predictions, providing practical insights even when direct ground truth evaluations are not available.

B. Assessing Dispersion

Figure displays RHALE summaries for two contrasting predictors:

- **Latitude** (x_2). The mean line slopes downward, revealing that properties located further north tend to be cheaper. The shaded ribbon and the vertical violin markers reveal wide scatter, indicating strong interaction with other coordinates such as longitude or housing age.
- **Median Income** (x_8). A pronounced upward relation is evident. Although variability exists, the standard-deviation band is noticeably narrower than in the latitude case, suggesting a more uniform contribution across the population.

These visual cues enable an analyst to separate robust trends from sections where predictions hinge on contextual variables.

C. Reliability of Adaptive Segmentation

To stress-test the interval-selection routine, the following protocol is applied:

1) Compute reference curves on the full training set using a dense equal partition of eighty slices. Because the sample is large, this proxy is treated as "truth".
2) Draw thirty random subsamples of size $1\,000$.
3) For each subsample, fit label=(c)
   a) fixed-width ALE with $K \in \{5, 10, 20, 40, 80\}$,
   b) RHALE with automatically chosen boundaries.
4) Evaluate mean absolute deviations $\mathcal{L}^{\mu}$ and $\mathcal{L}^{\sigma}$ between subsample estimates and the reference surface.

Figure reports the aggregated errors for the two highlighted predictors. The adaptive procedure consistently achieves or surpasses the best fixed-width setting in both mean and dispersion, while avoiding the trial-and-error process of picking K.

D. Summary of Findings

Real-data exploration supports three central messages:

1) *Mean trends alone are misleading*. Latitude exhibits a clear monotone decrease, yet the broad variability band warns that price sensitivity differs markedly across neighbourhoods. Income, by comparison, shows both a strong positive trend and lower heterogeneity.
2) *Automatic limits are dependable*. With one thousand points—only five percent of the full training sample—RHALE replicates reference curves with accuracy similar to, or better than, the most favourable fixed grid.
3) *No manual parameter tuning is required*. Analysts bypass the burden of grid-size selection; the dynamic programme allocates narrow slices where the derivative varies quickly and wider slices where the surface is flat.

VI. Conclusions and Prospects

RHALE enriches global explanation in two complementary ways. First, it augments traditional accumulated-effect curves with a quantitative spread measure, surfacing hidden

interactions and instance-specific behaviour. Second, it introduces a principled, data-driven partitioning rule that minimises approximation bias while keeping Monte-Carlo noise under control. Synthetic studies confirm superiority over PDP–ICE in correlated scenarios and over fixed-width ALE in diverse functional forms; an empirical case with housing prices illustrates interpretative advantages and computational efficiency.

Several avenues remain open:

- **Categorical inputs**. Extending the optimisation to mixed-type variables requires specialised distance measures and constraints.
- **High-dimensional targets**. Multi-output models call for joint dispersion metrics or task-specific aggregation rules.
- **Streaming environments**. Developing incremental updates would enable online monitoring of non-stationary systems without recomputing gradients from scratch.

Altogether, the proposed framework advances model-agnostic, information-rich explanation and lays groundwork for future research on adaptive, heterogeneous effect analysis.

REFERENCES

[1] C. Molnar, T. Freiesleben, G. König, G. Casalicchio, M. N. Wright, and B. Bischl, "Relating the partial dependence plot and permutation feature importance to the data generating process," *arXiv preprint arXiv:2109.01433*, 2021.

[2] J. Herbinger, B. Bischl, and G. Casalicchio, "Repid: Regional effect plots with implicit interaction detection," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022, pp. 10 209–10 233.

[3] M. Britton, "Vine: visualizing statistical interactions in black box models," *arXiv preprint arXiv:1904.00561*, 2019.

[4] J. H. Friedman and B. E. Popescu, "Predictive learning via rule ensembles," *The annals of applied statistics*, pp. 916–954, 2008.

[5] B. M. Greenwell, B. C. Boehmke, and A. J. McCarthy, "A simple and effective model-based variable importance measure," *arXiv preprint arXiv:1805.04755*, 2018.

[6] S. M. Lundberg, G. G. Erion, and S.-I. Lee, "Consistent individualized feature attribution for tree ensembles," *arXiv preprint arXiv:1802.03888*, 2018.

[7] R. K. Pace and R. Barry, "Sparse spatial autoregressions," *Statistics & Probability Letters*, vol. 33, no. 3, pp. 291–297, 1997.

Navigating Ethical Dilemmas in AI-Driven Supply Chain Operations

Harish Kasireddy
harishkasireddy87@gmail.com

***Abstract*—This paper explores the ethical dilemmas arising from the integration of Artificial Intelligence (AI) into supply chain operations. It examines critical issues such as data privacy, algorithmic bias, the impact on employment, and the need for transparency in AI decision-making. The analysis focuses on the challenges of ensuring responsible AI implementation in supply chain management to mitigate potential negative consequences and promote ethical practices.**

I. Introduction

A. Conceptualizing Artificial Intelligence in Supply Networks

Artificial Intelligence (AI) within supply networks denotes the deployment of advanced computational technologies—including machine learning (ML), natural language processing (NLP), robotics, and data-driven analytics—across operational workflows. The integration of these technologies seeks to elevate operational intelligence, enhance performance efficiency, and establish autonomous systems capable of self-optimization. In modern logistics environments, AI enables the interpretation of expansive datasets, generation of predictive insights, execution of automated decisions, and the creation of agile supply models that adjust dynamically to changing conditions.

B. Significance of AI in Enhancing Supply Chain Functionality

AI contributes transformative capabilities across distinct areas within supply chain ecosystems:

1) Transport and Routing: Advanced AI solutions contribute to transportation optimization through dynamic scheduling, adaptive routing, and shipment tracking. These systems factor in historical patterns, current traffic data, and environmental elements to recommend optimal delivery pathways, thus curtailing travel duration and fuel expenditure.

2) Demand Prediction: AI significantly refines demand forecasting, a critical aspect of stock regulation. By synthesizing previous transaction records, behavioral analytics, and external market indicators, machine learning models deliver precise and fluid projections. This reduces both overstock risks and inventory deficits by adapting to market variances.

3) Stock Oversight: Inventory control is enhanced through AI-driven predictive techniques, which evaluate inventory thresholds, forecast requirements, and facilitate automated restocking. Such models determine ideal quantity levels for various geographical nodes and support warehouse automation, lowering manual errors and accelerating order processing.

4) Vendor Analysis: AI mechanisms evaluate metrics such as vendor reliability, quality indices, and delivery timelines. This enables superior supplier selection and proactive disruption management, enhancing operational continuity and procurement integrity.

II. Ethical Implications of AI in Supply Chain Operations

Despite numerous advantages, the integration of AI introduces substantial ethical dilemmas:

A. Workforce Impact

Automation threatens to replace repetitive human roles including logistics planning and stock handling. This technological evolution poses challenges related to employment stability, especially in labor-intensive segments.

B. Information Protection and Confidentiality

Supply systems process sensitive information encompassing personal and commercial data. The application of AI raises concerns regarding data acquisition, usage boundaries, storage integrity, and exposure to digital threats.

C. Algorithmic Fairness

AI systems depend on training data, which may inherit historical societal prejudices. This can skew AI decisions in supplier assessment, demand analysis, or regional logistics, fostering inequity across systems.

D. System Transparency and Responsibility

Complex AI algorithms often operate opaquely, complicating error traceability and accountability. For instance, erroneous forecasting may result in service disruption, yet the source of fault—whether algorithmic, human, or systemic—can be difficult to isolate.

E. Ecological Considerations

The energy demand of AI infrastructure, particularly data centers, contributes to environmental degradation. Although AI supports sustainable practices such as route optimization and waste minimization, its environmental footprint must be critically evaluated.

III. Ethical Dimensions in AI Deployment

A. Bias Embedded in Algorithms

1) Procurement, Recruitment, and Logistics Decisions: Unbalanced training datasets can lead to partial outcomes. In vendor selection, legacy biases can influence preference algorithms. In hiring, skewed historical records may propagate demographic imbalances. In logistics, certain zones might face deprioritization based on biased delivery history.

2) Illustrations of Data Prejudice: Racial, gender, and socioeconomic biases in datasets can perpetuate discrimination. For example, preference for a specific demographic in historical hiring could result in algorithmic reinforcement of the same bias, distorting fairness in recruitment or vendor consideration.

3) Global Inequality Reinforcement: AI may inadvertently favor suppliers in technologically advanced regions, sidelining those from underdeveloped economies. Such bias can exacerbate global disparities by embedding socio-economic inequality into algorithmic processes.

B. Clarity and Ownership in AI Functionality

1) Demand for Transparency: Stakeholders must comprehend AI-driven conclusions. Lack of transparency can erode confidence and obscure the rationale behind key decisions, such as vendor prioritization or distribution reallocation.

2) Defining Accountability: In cases of operational failures due to AI miscalculations, clear accountability structures must be established. Whether the fault lies with the software engineers, system deployers, or dataset curators must be explicitly defined.

3) Explainable AI (XAI): XAI provides interpretability, allowing decision-makers to understand how outcomes were derived. This enhances trust and ensures alignment between AI outputs and organizational objectives.

C. Data Ethics and Privacy Safeguards

1) Handling of Sensitive Data: Personal and business-related data must be handled ethically. Compliance with data protection standards, such as respecting user consent and ensuring intended usage, is essential.

2) Security Vulnerabilities: AI systems may be prone to data infiltration. Breaches not only tarnish reputations but also jeopardize individual privacy, calling for fortified cybersecurity mechanisms.

3) Monitoring and Surveillance Implications: Monitoring tools powered by AI may infringe on privacy. Systems tracking employees or asset locations must be evaluated for their impact on worker autonomy and ethical labor conditions.

D. Labor Displacement and Human Rights

1) Automation-Induced Job Loss: AI solutions, though efficient, may displace human workers. Roles in transportation, warehousing, and order fulfillment are particularly susceptible, demanding human-centered automation strategies.

2) Exploitation via AI Oversight: AI-regulated workspaces might enforce unrealistic productivity benchmarks, inducing fatigue and compromising safety. Ethical governance must oversee AI's role in workforce monitoring.

3) Skills Transition Initiatives: Investment in skill development and retraining is necessary. Ethical AI integration must ensure that workers are equipped to occupy emerging roles within the digital economy.

E. Environmental Sustainability and AI

1) Enhancing Environmental Efficiency: AI systems aid in reducing waste, managing emissions, and promoting reuse through optimized routes and inventory. These capabilities contribute to eco-friendly operational models.

2) Energy Demand and Carbon Concerns: Training AI models requires significant energy. The sustainability benefits must be weighed against the ecological burden of running complex computational infrastructure.

3) Corporate Environmental Responsibility: Organizations must adopt environmental stewardship in AI deployment. Responsible practices should aim for long-term ecological viability over short-term profitability.

IV. Challenges in Deploying Artificial Intelligence in Supply Chain Management

A. Data Integrity and Accessibility

1) Complexity in Acquiring Reliable and Impartial Datasets: Artificial intelligence frameworks demand dependable and impartial datasets to function effectively. Nevertheless, obtaining structured, unbiased, and coherent data remains a core obstacle. Supply chain ecosystems collect information from disparate channels, including vendor data, shipment logs, and consumer transactions. When this information is inconsistent, erroneous, or partial, the accuracy of algorithmic outputs deteriorates, potentially impairing strategic outcomes. For instance, data contaminated with duplication or systemic prejudices can mislead predictive models, particularly in estimating market demand:

$$\widehat{D\_t} = \frac{1}{n} \sum \_i = 1^n \phi(x\_i) + \epsilon \tag{1}$$

where $\widehat{D\_t}$ denotes forecasted demand, $\phi(x\_i)$ represents transformed feature input, and ϵ is the error term from data distortion.

2) Fragmentation of Data Repositories: Operational silos within enterprises lead to compartmentalized data storage, often obstructing a unified analytical perspective. Procurement, logistics, and sales divisions may employ isolated databases, thereby hindering AI applications from interpreting a holistic view of operations. Additionally, cross-organizational data sharing is often limited due to privacy concerns, competitive barriers, or incompatible data formats, impeding AI systems from delivering synchronized insights across the supply chain network.

3) Expenditure and Logistical Challenges in Global Data Collection: Multinational supply chains encompass diverse actors across continents, complicating the aggregation of standardized and meaningful data. High financial costs and logistical burdens associated with harmonizing multilingual, multifaceted, and asynchronously updated records elevate the implementation barriers for AI across such networks.

B. Technological Integration Constraints

1) Incompatibility with Pre-existing Infrastructure: Organizations still reliant on outdated infrastructure—such as legacy Enterprise Resource Planning (ERP) platforms—often find integration of AI modules technologically restrictive. These legacy systems lack the scalability and interoperability required to process contemporary data-intensive tasks. As a result, aligning AI tools with such systems necessitates major infrastructural revisions.

2) Cultural Resistance from Organizational Stakeholders: Resistance to technological transformation is prevalent among personnel accustomed to manual workflows. Concerns surrounding job redundancy, unfamiliarity with intelligent tools, and skepticism from leadership regarding tangible returns from AI adoption present substantial socio-cultural barriers. This inertia delays digital transitions and demands robust change management protocols.

3) High Initial Investment for Integration: Incorporating AI into operational pipelines demands capital-intensive commitments for software procurement, staff training, system reconfiguration, and recruitment of data professionals. Although such investment promises future returns through optimized performance, the initial expenditure can be financially burdensome for mid-sized enterprises.

C. Security and Vulnerability Risks

1) Exposure to Digital Threats in Automated Systems: AI-enabled platforms are susceptible to cyber threats that can compromise critical supply chain functions. Threat actors may inject false data into algorithmic systems, distort inventory metrics, or redirect delivery mechanisms. Such disruptions can lead to major financial and operational setbacks:

$$C\_risk = \lambda(P\_a \cdot L\_i) \tag{2}$$

where C_risk is the calculated cyber risk cost, P_a denotes attack probability, and L_i reflects the loss impact.

2) Necessity of Algorithm Protection Mechanisms: To safeguard AI decision-making, robust defense mechanisms must be deployed to detect and prevent algorithmic manipulation. This includes implementing validation filters, adversarial training, and real-time anomaly detection to avoid deliberate input distortion.

3) Confidentiality of Proprietary Data Assets: Intellectual property (IP), proprietary models, and business-critical datasets constitute valuable assets that require encryption, access control, and secure cloud infrastructure to prevent unauthorized exposure or theft. Ensuring data integrity across partner networks is vital for maintaining competitive positioning.

D. Moral Responsibility in Automated Decisions

1) Maintaining Human Oversight for Strategic Judgments: While automation enhances speed and scalability, strategic decisions—particularly those involving social, environmental, or contractual nuances—require human arbitration. Autonomous systems should augment, not replace, ethical discernment in areas like supplier selection or community impact analysis.

2) Avoiding Over-Optimization at Ethical Expense: Optimization routines driven by cost or speed might inadvertently select vendors with poor labor standards or environmentally damaging practices. Therefore, ethical constraints must be embedded into optimization functions:

$$\min\_\mathbf{x}\, C(\mathbf{x}) \quad \text{s.t.} \quad E(\mathbf{x}) \leq \tau \tag{3}$$

Here, $C(\mathbf{x})$ is the cost function, and $E(\mathbf{x})$ denotes ethical impact constrained under a threshold τ.

3) Preserving Equitable Labor Conditions Amidst Automation: Automated workflows in warehousing and production facilities must not compromise labor dignity. There remains a critical obligation to ensure safe working conditions, fair compensation, and manageable workloads, even in highly automated environments.

V. Addressing Ethical Challenges and Mitigation Strategies

A. Legislation and Governance Frameworks

1) Establishment of Regulatory Infrastructures for AI Deployment in SCM: As artificial intelligence becomes increasingly integral to supply chain operations, the formulation of comprehensive and enforceable legal structures becomes imperative. These regulations would define acceptable AI applications and ensure alignment with ethical imperatives, encompassing protections for personnel, clientele, and ecological systems. Key domains include:

- **Data Governance Statutes:** Mandating adherence to data privacy statutes such as the GDPR and equivalent international frameworks to secure individual data during AI utilization.
- **Accountability Regulations:** Clarifying the attribution of responsibility when algorithmic errors result in operational mishaps such as inventory inaccuracies or delivery failures.
- **Operational Safety Standards:** Prescribing protocols for the secure integration of AI-based machinery within supply chain infrastructure to prevent harm to human operators.

2) Codification of Ethical Principles for AI-Enabled Enterprises: Beyond legislative mandates, normative standards are essential for steering the ethical deployment of AI in commercial operations. These benchmarks assist firms in embedding values such as equity, openness, and answerability in their AI initiatives. Areas of concern include:

- **Impartiality:** Guaranteeing algorithmic neutrality with respect to attributes such as ethnicity or gender.

- **Responsibility Attribution:** Designating individuals or roles accountable for AI outcomes.
- **System Interpretability:** Facilitating transparency in AI reasoning pathways to promote user confidence and oversight.

3) Global Entity Contributions to Ethical Standardization: Due to the transnational nature of supply networks, international agencies such as the ISO, OECD, and UN play pivotal roles in establishing and disseminating unified ethical AI frameworks. Their contributions include:

- Harmonizing AI conduct codes across jurisdictions.
- Mediating governmental cooperation to institute cross-border compliance mechanisms.
- Broadcasting validated ethical AI use cases to motivate widespread adoption.

B. Ethically Conscious AI Development

1) Advocating Team Inclusivity in AI Architecture: Incorporating a heterogeneous workforce in AI creation processes fosters the detection and rectification of systemic prejudices. Diverse input from varied cultural, racial, and gender perspectives enhances algorithmic fairness and ensures that AI tools cater equitably across societal segments.

2) Institutionalizing Accountability via Algorithm Audits: Routine verification of AI functionalities ensures ethical consistency and operational efficacy. Audit parameters may include:

- **Bias Identification:** Locating skewed decision patterns.
- **Performance Validation:** Confirming output accuracy under various operational contexts.
- **Decision Accountability:** Establishing audit trails to trace algorithm-driven outcomes.

3) Interdisciplinary Cooperation for Ethical Alignment: A synergistic approach involving AI developers, supply chain strategists, and ethics specialists fosters the integration of AI systems that satisfy both operational targets and societal expectations. Each contributor offers unique insights to mitigate unintended negative consequences and enhance benefit distribution.

C. Awareness and Training Initiatives

1) Capacity Building in Ethical AI Utilization: Supply chain personnel must be equipped with knowledge regarding AI's ethical dimensions. Instructional content should emphasize:

- Foundational ethical tenets such as justice, clarity, and responsibility.
- Possible detriments including labor displacement, surveillance misuse, and algorithmic bias.
- Practical guidance for ethical AI implementation and ongoing evaluation.

2) Highlighting Societal Impacts of AI Integration: Raising consciousness about the effects of AI on workforce dynamics and privacy rights is essential. Focus areas encompass:

- **Workforce Transformation:** Addressing the need for reskilling due to automation-induced role shifts.
- **Data Ethics:** Informing stakeholders on how personal data is handled and protected.

3) Institutionalizing Moral Decision-Making Protocols: Stakeholders should utilize ethical decision-making templates to direct AI-related actions. Important considerations involve:

- Ensuring algorithm decisions can be logically interpreted.
- Deploying strategies for mitigating prejudiced results.
- Assigning definitive responsibility for AI outcomes.

VI. Conclusion

A. Synthesis of Principal Ethical Considerations

Artificial intelligence has revolutionized supply chain management by enhancing operations in forecasting, procurement, logistics, and inventory oversight. Nonetheless, several moral concerns accompany its proliferation:

- **Training Data Bias:** Algorithms trained on non-representative datasets may foster discrimination in supplier or employee selection.
- **Opacity and Responsibility:** There is an urgent necessity to clarify culpability in instances of algorithmic failure.
- **Data Sovereignty and Cybersecurity:** Safeguarding vast datasets from unauthorized access is critical.
- **Employment Realignment:** Automation-driven displacement necessitates proactive upskilling and labor policy reform.
- **Environmental Footprint:** The energy demands of AI infrastructure require sustainable design and implementation strategies.

B. Imperative for Ethical Deployment Protocols

Organizations should adhere to structured AI adoption plans that incorporate:

- Comprehensive ethical doctrines encompassing justice and transparency.
- Collaboration among technical teams, supply professionals, and ethical analysts.
- Periodic system evaluations to adjust implementations in response to real-world implications.

C. Final Remarks on Responsible Innovation

It is essential to balance technological breakthroughs with fundamental human dignity and ecological stewardship. Ethical AI deployment ensures:

- Empowerment of labor through skill enrichment and equitable treatment.
- Resource-conscious operations aligned with climate goals.
- Holistic systems that serve stakeholders across all levels—individuals, enterprises, and the planet.

$$\text{AI\_ethical} = f(\text{Fairness}, \text{Transparency}, \text{Accountability}, \text{Sustainability})$$

Where AI_ethical denotes the net ethical benefit of AI deployment across supply chains, driven by fairness, openness, stakeholder accountability, and ecological awareness.

REFERENCES

[1] A.R. Teixeira, J.V. Ferreira, and A.L. Ramos, "Intelligent supply chain management: A systematic literature review on artificial intelligence contributions," _Information_, vol.16, no.5, art.399, 2025.

[2] F.S. "Examining the integration of artificial intelligence in supply chain," _Frontiers in AI_, Apr. 2024.

[3] A. Brintrup, "Trustworthy, responsible, ethical AI in manufacturing and supply chains: synthesis and emerging research questions," _arXiv_, May 2023.

[4] J. Cobbe, M.Veale, and J. Singh, "Understanding accountability in algorithmic supply chains," _arXiv_, Apr. 2023.

[5] M.A. Jahin, "AI in supply chain risk assessment: a systematic literature review and bibliometric analysis," _arXiv_, Dec. 2023.

[6] S. Nassar, "Towards a framework for responsible AI in supply chain management," in _Proc. 7th Int. Conf. ERPBSS_, Dubai, May 2024.

[7] D.Goswami, "A Systematic Literature Review on Artificial Intelligence Contributions," _Information_, vol.16, no.5, art.399, 2025.

[8] Author(s), "Integrating AI in sustainable supply chain practices: comparative analysis between the USA and Europe," _Int. J. of Computer Applications_, 2024.

[9] N. Author(s), "Taking a snapshot of artificial intelligence in supply chain," _ScienceDirect_, 2025.

[10] Author(s), "AI-Enabled supply chain management: a bibliometric analysis," _Sustainability_, vol.17, no.5, art.2092, 2025.

[11] Author(s), "Human–artificial intelligence collaboration in supply chain outcomes," _Ann. Oper. Res._, 2025.

[12] J. Mokander and L. Floridi, "Operationalising AI governance through ethics-based auditing: an industry case study," _arXiv_, Jul. 2024.

[13] Author(s), "The supply chain capitalism of AI: a call to rethink algorithmic …," _Information Polity_, 2024.

[14] Subharun Pal, "The Impact of AI on Global Supply Chain Management: A Review of Literature," _SEEJPH_, 2025.

[15] "Harnessing AI for agile and ethical supply chains: cost savings …," _WorldCertification.Org_, 2024.

[16] "Impact of artificial intelligence on supply chain optimization," _J. Tech. Stud._, 2024.