A publication of ISCA*:
International Society for Computers
and Their Applications

INTERNATIONAL JOURNAL OF COMPUTERS AND THEIR APPLICATIONS

TABLE OF CONTENTS

***"International Journal of Computers and Their Applications is Peer Reviewed".**

Volume 32, No. 1, March 2025 **ISSN 1076-5204**

International Journal of Computers and Their Applications

A publication of the International Society for Computers and Their Applications

Editorial

It is my distinct honor, pleasure, and privilege to serve as the Editor-in-Chief of the International Journal of Computers and Their Applications (IJCA) since 2022. I have a special passion for the International Society for Computers and their Applications. I have been a member of our society since 2014 and have served in various capacities. These have ranged from being on program committees of our conferences to being Program Chair of CATA since 2021 and currently serving as one of the Ex-Officio Board Members. I am very grateful to the ISCA Board of Directors for giving me this opportunity to serve society and the journal in this role.

I would also like to thank all the editorial board, editorial staff, and authors for their valuable contributions to the journal. Without everyone's help, the success of the journal would be impossible. I look forward to working with everyone in the coming years to maintain and further improve the journal's quality. I want to invite you to submit your quality work to the journal for consideration for publication. I also welcome proposals for special issues of the journal. If you have any suggestions to improve the journal, please feel free to contact me.

Dr. Ajay Bandi
School of Computer Science and Information Systems
Northwest Missouri State University
Maryville, MO 64468
Email: AJAY@nwmissouri.edu

In 2025, we are having four issues planned (March, June, September, and December). The next latest issue is taking shape with a collection of submitted papers.

I would also like to announce that I will begin searching for a few reviewers to add to our team. We want to strengthen our board in a few areas. If you would like to be considered, don't hesitate to get in touch with me via email with a cover letter and a copy of your CV.

Ajay Bandi, Editor-in-Chief
Email: AJAY@nwmissouri.edu

This issue of the International Journal of Computers and their Applications (IJCA) has gone through the normal review process. The papers in this issue cover a broad range of research interests in the community of computers and their applications.

IJCA Contributed Papers: This issue comprises papers that were contributed to the International Journal of Computers and their Applications (IJCA). The topics and main contributions of the papers are briefly summarized below:

Ngoc-Huan Le, Thanh-Hai Diep, Ngoc-Duc Trinh, Ngoc-Hay Nguyen, Viet-Thang Nguyen, Thanh-Son Nguyen and Narayan C. Debnath School of Engineering, Eastern International University Binh Duong Province, Viet Nam, Eastern International University, Binh Duong Province, VietNam present their work "**DEVELOPMENT OF A CYBER PHYSICAL SYSTEM FOR CONVENTIONAL MACHINES IN SMART FACTORIES**". In this study, the successful development of a Cyber-Physical System (CPS) tailored for four conventional machines in a real manufacturing environment was presented. Each machine was equipped with multiple sensors to monitor key operational parameters and ensure comprehensive data acquisition. The collected data is processed and visualized through an intuitive smart dashboard that can be accessed via a server computer and a web-based application. The proposed system allows for real-time monitoring, analysis, and report generation, including the automated calculation of Overall Equipment Effectiveness (OEE) and Overall Line Effectiveness (OLE) for operational efficiency. Besides, the CPS proactively identifies and mitigates potential errors, and enhances system reliability by implementing data thresholding techniques. Furthermore, the architecture can support predictive maintenance by analyzing trends and anomalies in sensor data. It paves the way for minimized downtime and cost savings. The CPS represents a significant advancement in digitizing conventional machines and manufacturing processes, contributing to increased efficiency, transparency, and scalability in line with the Industry 4.0 era.

Kalim Qureshi, Mohsen Al-Shamali, and Mostafa Abd-El-Barr. Kalim Qureshi, Mohsen Al-Shamali from Department of Information Science ,Kuwait University, Kuwait Former-Dean College of Computing Science and Engineering, Kuwait University, Kuwait, present their work "**An Empirical Study of Hardening Network Access Control Systems**". This article introduces the work presents a Network Access Control (NAC) is one of many solutions that plays a critical role in defining security policies in networking. Three open-source NAC solutions were analyzed and compared: OpenNAC, FreeNAC, and PacketFence. The results showed that the Packet Fence solution has better performance in terms of security features. Network layer-2 attacks were introduced against the candidate solution to verify vulnerabilities. These are Cisco Discovery Protocol, Dynamic Host Configuration Protocol, Spanning Tree Protocol, Dynamic Trunking Protocol, and VLAN Trunking Protocol. An enhanced PacketFence was proposed to mitigate network threats in a simulated environment; by using the network simulator tool (GNS3) and through hardening a critical component of PacketFence via applying supportive configurations and commands. We observed that the proposed enhancement solution improved network security. This is measured in terms of 22\% to 84\% increase in the CPU utilization during an attack that lasted for 10 minutes. In addition to root cost increase from 0 to 12 after launching 3 STP attacks. This is a substantial surge in MAC address table entries. Interface status was also changed to trunk and the VLAN entries were manipulated either by adding or removing entries in the VLAN table.

Nick Rahimi, , and Sindhuja Penchala from School of Computing Sciences & Computer Eng. University of Southern Mississippi Hattiesburg, US, present their work "**Maximizing Cyber Resilience through Efficient Vulnerability Prioritization: The WBTS Model**". This article introduces the in today's digital landscape, cybersecurity requires robust vulnerability management to mitigate potential threats effectively. Our study presents a risk prioritization approach that leverages weighted base scores and vulnerability titles to enhance threat assessment. This method enables organizations to systematically evaluate, classify, and prioritize vulnerabilities based on their impact, exploitability, and severity. By adopting this approach, security teams can allocate resources more efficiently, ensuring that the most critical threats are addressed promptly. The proposed model improves decision-making in cybersecurity strategies, reducing overall risk exposure. Furthermore, our approach aligns with industry best practices and regulatory requirements, reinforcing an organization's security posture against evolving cyber threats.

Antoine Bossard, from Graduate School of Science Kanagawa University, Yokohama, Kanagawa 221-8686, Japan, presented their work "**Image Processing Without Sacrificing the Functional Paradigm**". The paper discusses the Functional programming supports robust development, prioritizing the programmer over hardware and performance concerns. Features like map have influenced other paradigms, such as C++ and JavaScript. However, challenges remain, particularly with input-output (I/O) operations, which can deter users in memory-intensive applications. This paper demonstrates the practicality of functional programming for such scenarios using Racket (a Lisp dialect) for image processing. Both theoretical and experimental evaluations assess algorithm performance, including parallel processing on single and multiple cores.

Chaima BENSAID, Mohammed Khalil HADJ AHMED, Mohamed Mehdi BENALI from Computer Science Department, Khemis Meliana University, ALGERIA, presented their work "**Energy-Efficient Dynamic Cluster Formation for WSN Lifetime Optimization**". This study explores the Wireless Sensor Networks (WSNs) are widely used in environmental monitoring, industrial control, healthcare, and security, enabling intelligent data collection and surveillance. Each sensor node operates autonomously, equipped with sensors, processing units, and wireless communication modules. However, large-scale deployment poses challenges such as energy efficiency, communication reliability, fault tolerance, and security. This paper addresses these issues by optimizing the AOMDV routing protocol to enhance Quality of Service (QoS) and energy efficiency. Additionally, we propose a clustering algorithm that dynamically forms clusters based on node density and energy levels to reduce communication overhead and extend network lifetime. Simulations using NS2 demonstrate significant improvements in QoS, energy efficiency, and network longevity, offering promising solutions for WSN performance and sustainability.

SACI Abdallah and SEGHIR Rachid from Computer Science Department, University of BATNA 2, Algeria. Presented their work "**Optimizing Code Generation Efficiency Using the Polyhedral Model"** Optimizing scientific programs is crucial for performance, especially in resource-constrained environments like embedded systems and parallel computing. The polyhedral model has advanced affine-loop-nest code generation by improving parallelism and data locality. By transforming CLooG's mathematical representation, MPIB reduces costly function calls during loop traversal, improving execution time. Preliminary results show up to a 20% performance gain, particularly for larger parameter values, demonstrating its effectiveness over existing techniques.

Karim Morsi, Fatma najib, Wedad Hussein and Rasha Ismail from Faculty of computer and information science, Ain-Shams University Presented their work “**Arabic Text Summarization using transformer-based architectures”** Text summarizing is one of the most challenging tasks in natural language processing (NLP). This task is addressed in a large number of research projects and papers in the literature, but most of them focused on English language. Few studies are dealing with the complex Arabic language. Pre-trained Transformer-based language models have shown remarkable efficacy in addressing problems associated with text generation and natural language processing in recent times. However, there has not been much research on applying these models to Arabic text production. This study focuses on the implementation and fine-tuning pre-trained transformer-based language model structures for Arabic abstractive summarization, including AraBERT, mBERT models, and AraT5. We applied mBERT and AraBERT in the context of text summarization using a BERT2BERT-based encoder-decoder model. ROUGE measurements and manual human evaluation have been used to test the suggested models. Our models are trained and tested using XL-Sum Dataset of 46897 high-quality text-summary pairs. Their performance on out-of-domain data was also compared. We found that AraT5 outperforms AraBERT and mBERT Models, suggesting that a pre-trained Transformer with encoder-decoder functionality is more suited for text summarization. Moreover, AraT5 achieve high performance on out-of-domain dataset and received higher accuracy ratings in human evaluations compared to other models.

Galal eldin Abbas Eltayeb from Department of Management Information Systems, College of Business and Economics, Qassim University, Buraydah, Saudi Arabia Presented their work “**Data analytics enhance decision-making processes effectively using HoQ tool model**” Data analytics enhance decision-making by extracting valuable insights through data-based analysis. Tools like quality function deployment (QFD) simplify this process by aligning business strategies with customer needs. This paper proposes a QFD-based model to improve decision-making by defining strategic priorities and translating requirements into actionable improvements. The approach involves developing an application that implements a house of quality (HoQ) model, mapping enterprise requirements, conducting electronic surveys, and establishing evaluation criteria. A multi-attribute decision-making (MADM) process ranks alternatives, identifying the optimal enhancement strategy to support small manufacturing systems in adapting to market conditions.

As guest editors, we would like to express our deepest appreciation to the authors and the reviewers. We hope you will enjoy this issue of the IJCA. More information about ISCA society can be found at http://www.isca-hq.org.

Guest Editors:

Ajay Bandi, Northwest Missouri State University, USA.

March 2025

Development of A Cyber Physical System For Conventional Machines in Smart Factories

Ngoc-Huan Le[1], Thanh-Hai Diep[1], Ngoc-Duc Trinh[1], Ngoc-Hay Nguyen[1],
Viet-Thang Nguyen[1], Narayan C. Debnath[2], Thanh-Son Nguyen[1]
[1]School of Engineering, Eastern International University,
Binh Duong Province, Viet Nam
[2]School of Computing and Information Technology,
Eastern International University, Binh Duong Province, Viet Nam
Corresponding Author: huan.le@eiu.edu.vn

Abstract

In this study, the successful development of a Cyber-Physical System (CPS) tailored for four conventional machines in a real manufacturing environment was presented. Each machine was equipped with multiple sensors to monitor key operational parameters and ensure comprehensive data acquisition. The collected data is processed and visualized through an intuitive smart dashboard that can be accessed via a server computer and a web-based application. The proposed system allows for real-time monitoring, analysis, and report generation, including the automated calculation of Overall Equipment Effectiveness (OEE) and Overall Line Effectiveness (OLE) for operational efficiency. Besides, the CPS proactively identifies and mitigates potential errors, and enhances system reliability by implementing data thresholding techniques. Furthermore, the architecture can support predictive maintenance by analyzing trends and anomalies in sensor data. It paves the way for minimized downtime and cost savings. The CPS represents a significant advancement in digitizing conventional machines and manufacturing processes, contributing to increased efficiency, transparency, and scalability in line with the Industry 4.0 era.

Key Words: IoT; Cyber Physical System; Smart Factory; Industry 4.0; Digital Transformation; Digitalization; OEE.

1 Introduction

CPSs integrate the physical and digital worlds by embedding sensors, actuators, and software into industrial equipment, and allow precise monitoring and control. This integration fosters predictive maintenance that reduces downtime and optimizes resource utilization. By enabling seamless communication between devices, machines, and systems, the Internet of Things (IoT) plays a critical role in smart factories. Through IoT connectivity, real-time data is collected and shared across networks and helps enhance operational efficiency and decision-making. Together, IoT and CPS enable advanced automation, flexibility, and scalability, which are fundamental to Industry 4.0. By leveraging these technologies, smart factories can achieve unprecedented levels of productivity and innovation (Dornho¨fer et al. [10]; Averyanov et al. [2]). CPSs have found applications across a wide range of industries. In healthcare, CPS is utilized to enhance patient care and streamline medical procedures (Hemalatha et al. [13]; Rosado et al. [24]). Agriculture has benefited from CPS through innovations in precision farming and resource management (Hamzah et al. [12]). In transportation, CPS improves safety and efficiency by integrating intelligent systems(Wang and Liu [31]). Furthermore, CPS is a key enabler in smart city initiatives and drives sustainable urban development (Hemalatha et al. [14]). It also plays a role in water sustainability by monitoring and managing resources effectively (Cui [8]). CPS supports supply chain management and aids in the development of strategic policies across various sectors (Tonelli et al. [28]; Cheong and Lee [5]). Many studies have highlighted the vulnerability of CPS to external cyber-attacks, which can disrupt industries and lead to financial losses (Jamaludin and Rohani [17]). Besides, in (Oks et al. [23]), the authors present a novel categorization of industrial CPS across 10 sections, 32 areas, and 246 fields, and offer insights into future research directions to enhance Industry 4.0 applications. The others explore the defining characteristics, design methodologies, current state of the art, applications, challenges, and opportunities for addressing complex problems in the field of CPS (Lozano and Vijayan [21]). In (Habib and Chimsom I [11]), the authors highlight CPS's evolution toward intelligent, decision-making systems, their applications in smart cities, manufacturing, and supply chains, and the challenges of cybersecurity, real-time control, and interoperability that must be addressed for future advancements. The integration of digitization, Industry 4.0, the IoT, machine learning, and artificial intelligence is transforming the roles of plant operators and maintenance technicians. In (Wittenberg [33]), these advancements over the decades, current industry demands, and key research areas are examined. In ([1]), the authors investigate the interoperability challenges and integration of Digital Twins (DTs) within edge-enabled CPS in the context of Industry 4.0/5.0. The study identifies 77 interoperability challenges and proposes a framework with six levels—technical, syntactic, semantic, pragmatic, dynamic, and organizational—to help practitioners effectively adopt and use interconnected DTs in CPS. The CPSs Co-Simulator (CPS-Sim), a framework that integrates Matlab/Simulink for physical system simulation and QualNet (or OMNeT++) for communication network simulation is introduced. The key innovation lies in synchronizing these simulators with different time management methods, effectively demonstrated through a distributed clock synchronization algorithm in wireless sensor networks (Suzuki et al. [27]). Integrating CPS with IoT and Artificial Intelligence (AI) enables smart decision-making, and drives innovations in process optimization and customization. The researchers explore the integration of AI with CPS through Representation Learning (RepL) and emphasize its potential to extract meaningful abstractions from noisy sensor data and discrete system states. The study examines contemporary RepL methodologies applied to time-series data generated by CPS. A three-tank system as a case study to evaluate their strengths, limitations, and conditions for practical deployment in CPS contexts is used (Steude et al. [26]). The IoT holds a vital role in transforming traditional factories into smart factories in Industry 4.0. It enables predictive maintenance, energy optimization, and enhanced workplace safety through interconnected devices and sensors. Existing IoT connectivity solutions, highlighted IoT applications, technical challenges, and explored emerging technologies in smart factories are reviewed. They consist of predictive maintenance, asset tracking, inventory management, supply chain optimization, and so on (Ding et al. [9]; Soori et al. [25]; Cherif and Frikha [6]). In this paper, a CPS was successfully developed for two conventional milling machines and two conventional turning machines. The system collects and processes data from these machines, displaying it on a smart dashboard. Key performance metrics such as OEE and OLE are automatically calculated, providing valuable insights into operational efficiency. Four distinct Programmable Logic Controllers (PLCs) were integrated to manage the machines and ensure seamless data communication. By applying data thresholding techniques, the system effectively detects and prevents potential errors in order to enhance reliability. This CPS demonstrates a robust solution for digitizing conventional machines and aligning them with Industry 4.0 standards.

2 Literature Review

The field of industrial automation leverages CPS to optimize production processes and equipment performance. In (Hoffmann et al. [15]), the authors present a concept for the development, commercialization, operation, and maintenance of industrial CPSs in modern production, and highlight the challenges and opportunities for advancing both research and industrial practice. It defines the components and technological aspects of industrial CPS, compares them with traditional systems, and discusses key challenges and solutions to ensure the long-term sustainability of these systems. CPSs can transform technologies that bridge the physical and virtual worlds to create innovative applications and processes while dissolving traditional boundaries. The authors explore how CPS and IoT can drive a paradigm shift in manufacturing systems, optimize strategies, and introduce new applications, services, and data-driven business models (Kim and Park [19]). In (Chugh and Taqa [7]), the authors explore the role of Industry 4.0 technologies, including CPS and IoT, in transforming manufacturing through automation, real-time data exchange, and interconnected systems. By integrating physical components with software and communication networks, these technologies enhance decision-making, predictive maintenance, traceability, and production optimization. The proposed system is demonstrated through examples and a real-world case study. To increase productivity without the high costs of new machinery, the authors propose digitizing traditional machines by integrating motor controllers and sensors to collect and transmit data. This approach enhances the machining process by improving surface quality, reducing tool wear, and minimizing the risk of failure (Nguyen et al. [22]). In (Briatore and Braggio [3]), the research explores how Industry 4.0 technologies like IoT, DTs, and CPSs can revolutionize maintenance through predictive and prescriptive maintenance. By integrating these technologies into the Maintenance 4.0 framework, the study emphasizes resilience and environmental sustainability and proposes a six-step roadmap that begins with small-scale pilot projects to generate valuable results. The research explores the integration of Cloud Manufacturing and CPS through the use of OPC Unified Architecture (OPC UA) as a communication protocol to enable seamless data exchange and interoperability. The proposed hybrid architecture addresses challenges such as real-time monitoring, adaptive control, and efficient data management, and provides a pathway for optimizing manufacturing processes, and enhances real-time capabilities using cloud resources(Ji and Xu [18]). Besides, in (Va´squez-Capacho [30]), the authors introduce V-nets, a new formalism designed to address diagnosis challenges in CPS and industrial processes. V-nets are proposed as a reliable tool for managing fault detection and improving supervisory control in scenarios where traditional formal models of Discrete Event Systems (DES) fall short. The collaborative processes and model-based technologies used to develop a prototype Cyber-Physical Production System for USB sticks are detailed in (Zamfirescu and Neghina˘ [36]). It emphasizes co-simulation technology to enhance fidelity, enable independent subsystem validation, and facilitate structured dialogue between specialized teams. In the metallurgical industry, fused magnesia smelting for fused magnesium furnaces (FMF) is an energy-intensive process with high temperatures and complex dynamics. This process makes it challenging to measure and model the energy consumption per ton (ECPT) accurately. The paper introduces a CPS-based embedded optimal operational control system integrating advanced algorithms, industrial cloud computing, and wireless communication, successfully applied to ten FMF production lines in China. This integration significantly reduces the ECPT (you Chai et al. [35]). Cyber-physical production systems (CPPS), which link physical and digital components, serve as the backbone for these smart factories. It enables real-time management, adaptive processes, and optimization through global cooperation and innovation (Hozdic´ [16]). In (Torres et al. [29]), The authors focus on developing SmartBoxes using low-cost hardware like Raspberry Pi and industrial platforms such as NI CompactRIO, and employing OPC-UA and MQTT protocols for real-time data collection, processing, and integration. These SmartBoxes facilitate seamless interaction between supervisory systems and physical assets. And, a study presents a CPS-based thermal error compensator for CNC machine tools that are designed on an embedded system to rapidly collect sensor data, predict thermal errors, and communicate with CNC systems and cloud platforms (Lou et al. [20]). Applied to a CNC machine tool, the result demonstrates effective performance under various machining conditions. In the competitive manufacturing landscape, companies are integrating advanced technologies to enhance processes and productivity and align with Industry 4.0 principles. The research examines the transformation process of a factory producing spherical bushels. They utilize FlexSim software to create a production simulation platform for real-time management of production, supply, and logistics via Material Requirement Planning (MRP) and CPPS. The simulation optimized by using a load-capacity adjustment method. It improves equipment occupancy rates, demonstrates significant efficiency gains, and lays the groundwork for a future digital twin of the company (Chakroun et al. [4]). In (Williams et al. [32]), the authors present the implementation of DT of Cyber-Physiscal Tormach CNC machines, which replicates physical manufacturing operations by generating tool path positional values along the X, Y, and Z axes. The DT uses the MTConnect communication protocol to collect and store data in standardized XML and JSON formats for analysis. Validation was carried out by simulating and manufacturing a coin geometry on a real CNC machine. The results show a high correlation between the DT and real system. By integrating supervisory control and data acquisition (SCADA), edge computing, and cloud computing to monitor and analyze data streams from CNC machines and sensors, in (Yang et al. [34]) the authors propose a new data analysis framework for CPS. The framework employs signal smoothing and anomaly pattern detection techniques to identify and store significant patterns in the data stream. These patterns can then be used for further analysis and applications within CPS.

3 Research Methodology

In this project, four new control boxes for four machines were built. Each machine was equipped with a PLC controller from a different brand because, in reality, the company is using a wide variety of PLCs from different manufacturers and generations. A central server simultaneously connects to all four PLCs to collect data and store it in a shared database. A smart dashboard to show critical values that need to be monitored and provide alerts when sensor signals indicate that the equipment is about to operate abnormally was developed. Two important indicators calculated and displayed on the dashboard are OEE and OLE. Among them, Overall Equipment Efficiency (OEE) is a critical indicator, calculated as follows:

$$OEE = A \text{ x } P \text{ x } Q \quad (1)$$

where: A = Run Time / Planned Production Time. This measures how much time the production line was running as planned.
P = (Ideal Cycle Time × Total Count) / Run Time. This assesses whether the line is running at its maximum speed or capacity.
Q = Good Count / Total Count. This evaluates the proportion of defect-free products.

Planned Production Time and Ideal Cycle Time are predefined values that can be set in advance. Total Count represents the total number of products produced. Good Count refers to the quantity of products meeting quality standards. Both Total Count and Good Count are manually recorded. Consequently, Run Time is the only variable that requires automatic measurement during production. To address this, the authors propose capturing this value by transmitting an on/off signal to the data center whenever the machine starts or stops. With accurate information on machine start and stop times, Run Time can be calculated efficiently and reliably. Besides, OLE is a metric used to evaluate the performance and efficiency of an entire production line. It is similar to OEE but focuses on the performance of multiple machines or stations working together in a line. OLE was also calculated and displayed in this project. Based on actual production needs, three parameters (Capacity Utilization, First Pass Yield, and Scrap Rate) are also calculated using formulas (2), (3), and (4) and displayed. Where:

Capacity Utilization = Actual output / Maximum possible output (2)

First pass yield = (the number of units successfully produced without rework) / (the total number of units entering the process) (3)

Scrap rate = the amount of scrap / the total amount of output (4)

By equipping inverters, display screens, and sensors, the two traditional milling machines have been digitized as follows:

- The system for controlling spindle speed in milling machines has been enhanced by integrating a 3-phase inverter. It allows precise speed adjustments for optimal machining performance, product quality, and tool life. Additionally, the authors propose using optical sensors to record spindle speed data with high accuracy and enable analysis to identify and address factors affecting tool life and surface quality.

- Monitoring coolant parameters is essential for efficient milling machine operations. The coolant helps manage temperature, clear abrasive particles, and maintain machining quality. By using ultrasonic and thermal sensors, the developed CPS can automate the collection of coolant level and temperature data, enable better maintenance planning, and address issues such as low density, insufficient coolant levels, or overheating.
- In production, electric energy consumption during system operation is a crucial factor that can be calculated using wattage and spindle torque. Through torque value inspection, the operator can assess the compatibility between the feed rate, spindle speed, and the material properties of the workpiece. This process helps prevent tool breakage or wear during machining and ensures accuracy by enabling corrective actions for tool wear. Similarly, the digitization of two traditional lathes integrates sensors, control systems, and visualization tools with the following enhancements:
- Spindle speed monitoring: an optical sensor has been installed to record high-precision spindle speed data for real-time analysis. This enables optimization of machining processes and enhances tool life.
- Temperature monitoring: two temperature sensors measure spindle and motor temperature that provide critical insights to prevent overheating, ensure operational stability, and help condition-based maintenance planning.
- Energy efficiency monitoring: a current sensor is utilized to monitor the machine's power consumption that supports efficient energy management and early fault detection through monitoring the abnormal changes in current.
- Precision control system: a three-phase variable frequency drive (VFD) replaced traditional contactor-based motor control to allow step-less control and precise speed adjustments for enhancing machining accuracy, product quality, and tool life. A three-tier light tower provides real-time operational status updates, improves situational awareness, and ensures machine safety.
- Monitoring operations and observing graphical data: A Human-Machine Interface (HMI) has been installed to provide intuitive control and monitor operational parameters that can improve productivity and interaction between operators and machines.

This digitization establishes a foundation for cyber-physical integration in advanced manufacturing systems. Based on practical working experience, key indicators that directly affect product quality, machine failure, and tool breakage—such as spindle speed and coolant temperature—have been set with thresholds to notify operators before issues occur. Critical data is stored and analyzed, and reports can be easily generated to facilitate the management process. To easily monitor the dashboard in the areas inside and outside the company, a web-based application with a control flow as shown in Figure 1 was developed.

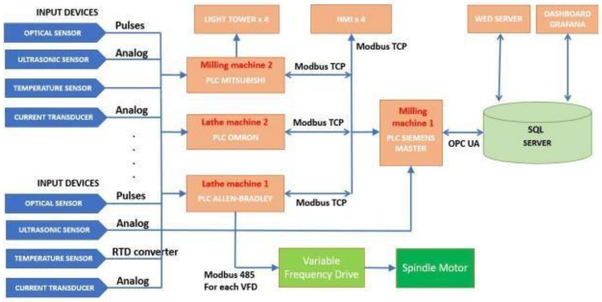


Figure 1: Connection diagram and transmission protocol of devices in the CPS system.

The diagram illustrates a system where various input devices, such as optical sensors, ultrasonic sensors, temperature sensors, and current transducers, provide essential data to multiple PLCs, including Siemens, Mitsubishi, Omron, and Allen-Bradley. Each sensor serves a specific purpose: optical sensors measure the spindle motor's speed, ultrasonic sensors monitor the coolant water level, temperature sensors track both the spindle motor's temperature and the coolant water's temperature, and current transducers measure the total current consumption of the machine. The Siemens PLC acts as the master, consolidating data from other stations and transmitting it to an SQL server using OPC UA for further processing. This data supports web-based monitoring and dashboard visualization via Grafana. Additionally, at each station, the PLCs control spindle motors using Variable Frequency Drives (VFD) via Modbus 485, with all configurations and operations at each machine being managed through the HMI. Furthermore, the spindle motor's torque is calculated based on the current measured from the VFD to enable precise motor management and system optimization. This setup integrates data acquisition, control, and visualization for efficient industrial automation.

4 Results and Discussion

Figure 2 depicts the actual CPS system with four traditional machines (02 milling and 02 turning) connected and transmitting data to the server computer. The inverters, sensors, light towers, and control cabinets integrated into each machine are all operating as planned. The entire system was operated simultaneously to test connectivity and the seamless flow of data. Additionally, to ensure the accuracy of OEE and OLE values, the CPS was continuously operated for several days.

For the milling machines, four sensors are installed at suitable locations as below:

- An optical sensor is mounted on the gearbox of the main motor to measure spindle speed.
- A temperature sensor is positioned in the coolant reservoir to monitor coolant temperature.
- An ultrasonic sensor is also installed in the coolant reservoir to measure the coolant level.
- A current sensor is placed on the power line within the machine's electrical cabinet to track its current consumption.

Similarly, the digitization of two traditional lathes includes the installation of four sensors, control systems, and visualization tools with the configuration as follows:

- **Spindle speed monitoring:** An optical sensor is installed inside the headstock to record high-precision spindle rotational speed data for real-time analysis.
- **Spindle motor temperature monitoring:** A temperature sensor is mounted on the spindle motor housing to measure its temperature.
- **Spindle temperature monitoring:** Another temperature sensor is installed within the headstock to measure the spindle temperature.
- **Energy efficiency monitoring:** A current sensor is located in the electrical cabinet to track the lathe's power consumption and energy usage.

Additionally, a three-color light tower is mounted at the top of the machines to indicate its operational status. The data collected from these sensors is processed and displayed on the HMI which is mounted on the control panel. This HMI interface allows operators to monitor machine parameters and adjust control settings flexibly. The spindle speed is managed by a VFD located in the electrical cabinet and can be modified either through the HMI or the dashboard. This digitization establishes a foundation for cyber-physical integration in advanced manufacturing systems.

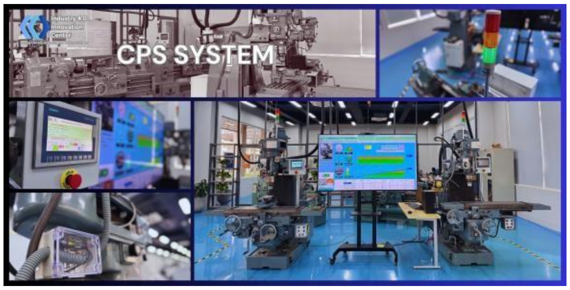
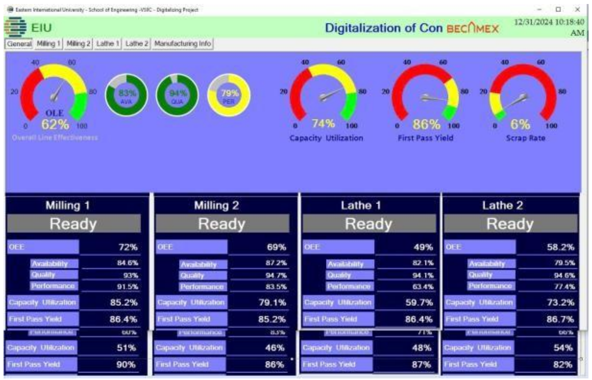


Figure 2: CPS system.

A professionally designed smart dashboard includes a general page summarizing the statistics of all four machines (Figure 3). On the General Page, which displays the overall statistics for all machines, the top part features graphs of the statistical values: OLE, Capacity Utilization, First Pass Yield, and Scrap Rate. The bottom part shows the key values of these parameters for each machine. Additionally, the OLE value of the whole system consisting of 4 machines is also calculated and displayed.



	Milling 1	Milling 2	Lathe 1	Lathe 2
	Ready	Ready	Ready	Ready
OEE	72%	69%	49%	58.2%
Availability	84.6%	87.2%	82.1%	79.5%
Quality	93%	94.7%	94.1%	94.6%
Performance	91.5%	83.5%	63.4%	77.4%
Capacity Utilization	85.2%	79.1%	59.7%	73.2%
First Pass Yield	86.4%	85.2%	86.4%	86.7%
Capacity Utilization	51%	46%	48%	54%
First Pass Yield	90%	86%	87%	82%

Figure 3: General page – Smart Dashboard on the server machine.

Each machine also has its own dedicated page displaying data from sensors, statistics, and warnings to alert operators about potential errors (Figure 4). The current basic awareness consists of Excessively High Coolant Temperature, Low Coolant Level, Abnormal Spindle Torque, and Excessively High Spindle Temperature. The displayed data for each machine are Machine operating status (Run/Idle/Error), OEE, Name of operator, Sensor data, power consumption (in watts and currency), warning messages, and statistics of specific data. On the left side of the interface, the machine's image is displayed along with the manufacturer's name and model. Below this, the operator's information is shown, including the name and photo. The system tracks data about the primary operator, the machine they are using, and their working hours. This information helps to visually identify the current operator and review operation history. The system includes a feature to change the operator by double-clicking the "Change Operator" option. This allows another operator to use the machine with permission and records each operator's operating time. As a result, any issues caused by operators can be traced back through the stored database. All data collected from the 4 machines are stored and can be exported according to the storage time when needed.

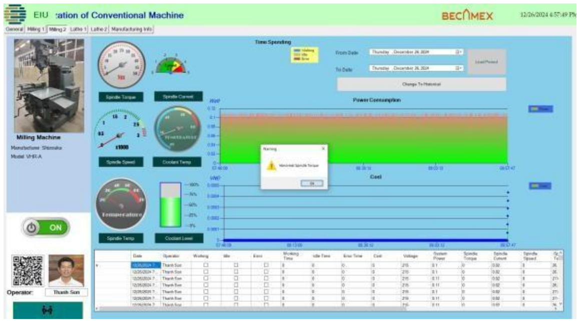


Figure 4: The user interface of the milling machine with the warning message "Abnormal Spindle Torque"

The web-based interface of the system is displayed as shown in Figure 5. Through this interface on the web-based application, operators can remotely monitor the operational status of the entire CPS system. A mobile application is currently being developed to facilitate monitoring.

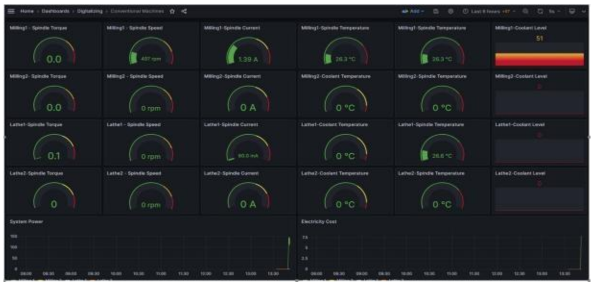


Figure 5: Smart Dashboard on the web application.

The result of the project is a favorable first step toward developing predictive maintenance modes as well as creating an automated production plan to optimize the production planning process. This is a basic, typical CPS system that includes the following functions:

- **Sensors:** Collect data from the physical environment.
- **Control Systems:** Process data collected from sensors and make decisions or commands to control physical devices.
- **Communication Network:** Connect sensors, control devices, and computers or servers to transmit data between components in the system. This network may include both wireless and wired protocols.
- **Computing and Software Processing:** Process and analyze data from sensors, and provide functions such as real-time analysis, forecasting, optimization, and automatic decision-making.
- **Actuators and Physical Devices:** Execute physical actions based on control commands from the system.
- **User Interface:** Allow operators to monitor, supervise, and control the system.
- **Safety:** Prevent damage or accidents.
- **Data Integration and Analytics:** Integrate data from various sources and use analytical algorithms to extract useful information, optimize system performance, and support decision-making.

5 Conclusions

In this paper, the successful development of a CPS designed to integrate four conventional machining tools is presented. The system effectively collects and visualizes data through a smart dashboard, and generates reports and alerts for potential risks that could impact the equipment. The alerts are based on data analysis from various sensors to ensure timely responses to anomalies. The CPS not only enhances real-time monitoring and operational reliability but also serves as a foundational step toward transforming traditional manufacturing systems into smart factories. The proposed system is full of functions that a standard CPS needs. By adopting this system, manufacturers can align with Industry 4.0 objectives so that they can achieve higher efficiency, improved decision-making, and greater adaptability to the demands of modern production environments.

References

[1] Sarthak Acharya, Arif Ali Khan, and Tero Pa¨iva¨rinta. Interoperability levels and challenges of digital twins in cyber–physical systems. *Journal of Industrial Information Integration*, 42:100714, 2024. ISSN 2452-414X. doi: https://doi.org/10.1016/j.jii.2024.100714. URL https://www.sciencedirect.com/science/article/pii/S2452414X24001572.

[2] Anton Averyanov, Shohin Aheleroff, Jan Polzer, and Xun Xu. Digitising a machine tool for smart factories. *Machines*, 10(11), 2022. ISSN 2075-1702. doi: 10.3390/machines10111093. URL https://www.mdpi.com/2075-1702/10/11/1093.

[3] F. Briatore and M. Braggio. Resilience and sustainability plants improvement through maintenance 4.0: Iot, digital twin and cps framework and implementation roadmap. *IFAC-PapersOnLine*, 58(8):365–370, 2024. ISSN 2405-8963. doi: https://doi.org/10.1016/j.ifacol.2024.08.148. URL https://www.sciencedirect.com/science/article/pii/S240589632400870X. 6th IFAC Workshop on Advanced Maintenance Engineering, Services and Technology AMEST 2024.

[4] Ayoub Chakroun, Yasmina Hani, Abderrahmane Elmhamedi, and Faouzi Masmoudi. Digital transformation process of a mechanical parts production workshop to fulfil the requirements of industry 4.0. In *2022 14th International Colloquium of Logistics and Supply Chain Management (LOGISTIQUA)*, pages 1–6, 2022. doi: 10.1109/LOGISTIQUA55056.2022.9938099.

[5] Hee-Woon Cheong and Hwally Lee. Technology and policy strategies in the era of cps (cyber physical system) and automated driving. *Procedia Computer Science*, 122:102–105, 2017. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2017.11.347. URL https://www.sciencedirect.com/science/article/pii/S1877050917325760. 5th International Conference on Information Technology and Quantitative Management, ITQM 2017.

[6] Mouna Regaieg Cherif and Hela Moalla Frikha. Iot device connectivity selection using a hybrid multi-criteria group decision-making approach. In *2023 International Conference on Decision Aid Sciences and Applications*

(DASA), pages 436–440, 2023. doi: 10.1109/DASA59624. 2023.10286610.

[7] Jimmy Chugh and Amer Taqa. Cyber-physical system (cps) & internet of things (iot) in manufacturing. 8:2319–5045, 11 2019.

[8] Xinyue Cui. Cyber-physical system (cps) architecture for real-time water sustainability management in manufacturing industry. *Procedia CIRP*, 99:543–548, 2021. ISSN 2212-8271. doi: https://doi.org/10.1016/j.procir.2021.03.074. URL https://www.sciencedirect.com/science/article/pii/S2212827121003619. 14th CIRP Conference on Intelligent Computation in Manufacturing Engineering, 15-17 July 2020.

[9] Jie Ding, Mahyar Nemati, Chathurika Ranaweera, and Jinho Choi. Iot connectivity technologies and applications: A survey. *IEEE Access*, 8:67646–67673, 2020. doi: 10.1109/ACCESS.2020.2985932.

[10] Mareike Dornho¨fer, Simon Sack, Johannes Zenkert, and Madjid Fathi. Simulation of smart factory processes applying multi-agent-systems—a knowledge management perspective. *Journal of Manufacturing and Materials Processing*, 4(3), 2020. ISSN 2504-4494. doi: 10. 3390/jmmp4030089. URL https://www.mdpi.com/2504-4494/4/3/89.

[11] Maki K. Habib and Chukwuemeka Chimsom I. Cps: Role, characteristics, architectures and future potentials. *Procedia Computer Science*, 200:1347–1358, 2022. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2022. 01.336. URL https://www.sciencedirect.com/science/article/pii/S1877050922003453. 3rd International Conference on Industry 4.0 and Smart Manufacturing.

[12] Muzaffar Hamzah, Md. Monirul Islam, Shahriar Hassan, Md. Nasim Akhtar, Most. Jannatul Ferdous, Muhammed Basheer Jasser, and Ali Wagdy Mohamed. Distributed control of cyber physical system on various domains: A critical review. *Systems*, 11(4), 2023. ISSN 2079-8954. doi: 10.3390/systems11040208. URL https://www.mdpi.com/2079-8954/11/4/208.

[13] T. Hemalatha, A. Bhuvaneswari, N. Poornima, B. Shubha, K. Santhi, M. Lawanyashri, and Geeta C. Mara. Secure and private data sharing in cps e-health systems based on cb-smo techniques. *Measurement: Sensors*, 27:100787, 2023. ISSN 2665-9174. doi: https://doi.org/10.1016/j.measen.2023.100787. URL https://www.sciencedirect.com/science/article/pii/S266591742300123X.

[14] T. Hemalatha, K. Sangeetha, K. Sasi Kala Rani, K.V. Kanimozhi, M. Lawanyashri, K. Santhi, and R. Deepalakshmi. Cps in block chain smart city application based on distributed ledger based decentralized technique. *Measurement: Sensors*, 30:100906, 2023. ISSN 2665-9174. doi: https://doi.org/10.1016/j.measen. 2023.100906. URL https://www.sciencedirect.com/science/article/pii/S2665917423002428.

[15] Martin W. Hoffmann, Somayeh Malakuti, Sten Gru¨ner, Soeren Finster, Jo¨rg Gebhardt, Ruomu Tan, Thorsten Schindler, and Thomas Gamer. Developing industrial cps: A multi-disciplinary challenge. *Sensors*, 21(6), 2021. ISSN 1424-8220. doi: 10.3390/s21061991. URL https://www.mdpi.com/1424-8220/21/6/1991.

[16] Elvis Hozdic´. Smart factory for industry 4.0: A review. *Journal of Modern Manufacturing Systems and Technology*, 7:28–35, 01 2015.

[17] Juliza Jamaludin and Jemmy Mohd Rohani. Cyber-physical system (cps): State of the art. In *2018 International Conference on Computing, Electronic and Electrical Engineering (ICE Cube)*, pages 1–5, 2018. doi: 10.1109/ICECUBE.2018.8610996.

[18] Tang Ji and Xun Xu. Exploring the integration of cloud manufacturing and cyber-physical systems in the era of industry 4.0 – an opc ua approach. *Robotics and Computer-Integrated Manufacturing*, 93:102927, 2025. ISSN 0736-5845. doi: https://doi.org/10.1016/j.rcim. 2024.102927. URL https://www.sciencedirect.com/science/article/pii/S073658452400214X.

[19] SungHyun Kim and Sungbum Park. Cps(cyber physical system) based manufacturing system optimization. *Procedia Computer Science*, 122:518–524, 2017. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2017.11.401. URL https://www.sciencedirect.com/science/article/pii/S1877050917326467. 5th International Conference on Information Technology and Quantitative Management, ITQM 2017.

[20] Ping Lou, Wenfeng Li, Jianmin Hu, Jiwei Hu, and Angran Xiao. Design and development of a thermal error compensator based on cps for cnc machine tools. pages 1–5, 03 2018. doi: 10.1109/ICNSC.2018.8361353.

[21] Carolina Villarreal Lozano and Kavin Kathiresh Vijayan. Literature review on cyber physical systems design. *Procedia Manufacturing*, 45:295–300, 2020. ISSN 2351-9789. doi: https://doi.org/10.1016/j.promfg.2020.04.020. URL https://www.sciencedirect.com/science/article/pii/S2351978920310581. Learning Factories across the value chain – from innovation to service – The 10th Conference on Learning Factories 2020.

[22] Thang Nguyen, Thanh-Hai Diep, THANH-SON NGUYEN, NGOC-DUC TRINH, NGOC-HUAN LE, and

NGOC-HAY NGUYEN. Digitalizing conventional machines: A complete practical implementation solution. *MM Science Journal*, 2023, 10 2023. doi: 10.17973/MMSJ.2023_10_2023056.

[23] Sascha Julian Oks, Max Jalowski, Michael Lechner, Stefan Mirschberger, Marion Merklein, Birgit Vogel-Heuser, and Kathrin M. Mo¨slein. Cyber-physical systems in the context of industry 4.0: A review, categorization and outlook. *Inf. Syst. Frontiers*, 26:1731–1772, 2022. URL https://api.semanticscholar.org/CorpusID:248036052.

[24] David G. Rosado, Antonio Santos-Olmo, Luis Enrique Sa´nchez, Manuel A. Serrano, Carlos Blanco, Haralambos Mouratidis, and Eduardo Ferna´ndez-Medina. Managing cybersecurity risks of cyber-physical systems: The marisma-cps pattern. *Computers in Industry*, 142:103715, 2022. ISSN 0166-3615. doi: https://doi.org/10.1016/j.compind.2022.103715. URL https://www.sciencedirect.com/science/article/pii/S0166361522001129.

[25] Mohsen Soori, Behrooz Arezoo, and Roza Dastres. Internet of things for smart factories in industry 4.0, a review. *Internet of Things and Cyber-Physical Systems*, 3:192–204, 2023. ISSN 2667-3452. doi: https://doi.org/10.1016/j.iotcps.2023.04.006. URL https://www.sciencedirect.com/science/article/pii/S2667345223000275.

[26] Henrik Steude, Alexander Windmann, and Oliver Niggemann. Learning physical concepts in cps: A case study with a three-tank system. *IFAC-PapersOnLine*, 55(6):15–22, 2022. ISSN 2405-8963. doi: https://doi.org/10.1016/j.ifacol.2022.07.099. URL https://www.sciencedirect.com/science/article/pii/S2405896322004840. 11th IFAC Symposium on Fault Detection, Supervision and Safety for Technical Processes SAFEPROCESS 2022.

[27] Atsushi Suzuki, Kazuyuki Masutomi, Isao Ono, Hideaki Ishii, and Takashi Onoda. Cps-sim: Co-simulation for cyber-physical systems with accurate time synchronization. *IFAC-PapersOnLine*, 51(23):70–75, 2018. ISSN 2405-8963. doi: https://doi.org/10.1016/j.ifacol.2018.12.013. URL https://www.sciencedirect.com/science/article/pii/S2405896318335456. 7th IFAC Workshop on Distributed Estimation and Control in Networked Systems NECSYS 2018.

[28] Flavio Tonelli, Melissa Demartini, Massimo Pacella, and Roberta Lala. Cyber-physical systems (cps) in supply chain management: from foundations to practical implementation. *Procedia CIRP*, 99:598–603, 2021. ISSN 2212-8271. doi: https://doi.org/10.1016/j.procir.2021.03.080. URL https://www.sciencedirect.com/science/article/pii/S221282712100370X. 14th CIRP Conference on Intelligent Computation in Manufacturing Engineering, 15-17 July 2020.

[29] Pedro Torres, Roge´rio Dion´ısio, Se´rgio Malha˜o, Lu´ıs Neto, Ricardo Ferreira, Helena Gouveia, and He´lder Castro. Cyber-physical production systems supported by intelligent devices (smartboxes) for industrial processes digitalization. In *2019 IEEE 5th International forum on Research and Technology for Society and Industry (RTSI)*, pages 73–78, 2019. doi: 10.1109/RTSI.2019.8895553.

[30] J.W. Va´squez-Capacho. V-nets, new formalism to manage diagnosis problems in cyber-physical systems (cps) and industrial applications. *IFAC-PapersOnLine*, 53(5):197–202, 2020. ISSN 2405-8963. doi: https://doi.org/10.1016/j.ifacol.2021.04.224. URL https://www.sciencedirect.com/science/article/pii/S2405896321004080. 3rd IFAC Workshop on Cyber-Physical & Human Systems CPHS 2020.

[31] Zezhou Wang and Xiang Liu. Cyber security of railway cyber-physical system (cps) – a risk management methodology. *Communications in Transportation Research*, 2:100078, 2022. ISSN 2772-4247. doi: https://doi.org/10.1016/j.commtr.2022.100078. URL https://www.sciencedirect.com/science/article/pii/S2772424722000282.

[32] Bethanie Williams, Gabriela Ciocarlie, Kyle Saleeby, Muhammad Ismail, and Clifton Mulkey. Digital twin of cyber-physical cnc for smart manufacturing. In *2023 IEEE 3rd International Conference on Digital Twins and Parallel Intelligence (DTPI)*, pages 1–6, 2023. doi: 10.1109/DTPI59677.2023.10365463.

[33] Carsten Wittenberg. Challenges for the human-machine interaction in times of digitization, cps & iiot, and artificial intelligence in production systems. *IFAC-PapersOnLine*, 55(29):114–119, 2022. ISSN 2405-8963. doi: https://doi.org/10.1016/j.ifacol.2022.10.241. URL https://www.sciencedirect.com/science/article/pii/S2405896322022686. 15th IFAC Symposium on Analysis, Design and Evaluation of Human Machine Systems HMS 2022.

[34] Chao-Lung Yang, Hendri Sutrisno, Nai-Wei Lo, Zhi-Xuan Chen, Ching-Chih Wei, Han-Wei Zhang, Chin-Teng Lin, Chen-Lung Wei, and Shang-Heng Hsieh. Streaming data analysis framework for cyber-physical system of metal machining processes. In *2018 IEEE Industrial Cyber-Physical Systems (ICPS)*, pages 546–551, 2018. doi: 10.1109/ICPHYS.2018.8390764.

[35] Tian you Chai, Zhi wei Wu, and Hong Wang. A cps based optimal operational control system for fused magnesium furnace. *IFAC-PapersOnLine*, 50 (1):14992–14999, 2017. ISSN 2405-8963. doi: https://doi.org/10.1016/j.ifacol.2017.08.2566. URL https://www.sciencedirect.com/science/article/pii/S2405896317334936. 20th IFAC World Congress.

[36] Constantin-Ba˘la˘ Zamfirescu and Mihai Neghina˘. Collaborative development of a cps-based production system. *Procedia Computer Science*, 162:579–586, 2019. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2019.12.026. URL https://www.sciencedirect.com/science/article/pii/S1877050919320368. 7th International Conference on Information Technology and Quantitative Management (ITQM 2019): Information technology and quantitative management based on Artificial Intelligence.

An Empirical Study of Hardening Network Access Control Systems

Kalim Qureshi*
Department of Information Science ,Kuwait University, Kuwait
Mohsen Al-Shamali †
Department of Information Science ,Kuwait University, Kuwait.
Mostafa Abd-El-Barr‡
Former-Dean College of Computing Science and Engineering, Kuwait University, Kuwait.

Abstract

Network Access Control (NAC) is one of many solutions that plays a critical role in defining security policies in networking. Three open-source NAC solutions were analyzed and compared: OpenNAC, FreeNAC, and PacketFence. The results showed that the PacketFence solution has better performance in terms of security features. Network layer-2 attacks were introduced against the candidate solution to verify vulnerabilities. These are Cisco Discovery Protocol, Dynamic Host Configuration Protocol, Spanning Tree Protocol, Dynamic Trunking Protocol, and VLAN Trunking Protocol. An enhanced PacketFence was proposed to mitigate network threats in a simulated environment; by using the network simulator tool (GNS3) and through hardening a critical component of PacketFence via applying supportive configurations and commands. We observed that the proposed enhancement solution improved network security. This is measured in terms of 22% to 84% increase in the CPU utilization during an attack that lasted for 10 minutes. In addition to root cost increase from 0 to 12 after launching 3 STP attacks. This is a substantial surge in MAC address table entries. Interface status was also changed to trunk and the VLAN entries were manipulated either by adding or removing entries in the VLAN table.

Key Words:Network Access Control (NAC); Network Security; PacketFence; Policy Enforcement Point; Hardening Configuration; Security performance improvements.

1 Introduction

The demands for Information Security has recently increased to a level that demanded every firm to have a dedicated team responsible for identifying information vulnerabilities to mitigate diverse threats encountered. Access control is a security technique that is used to organize the accessibility of assets and resources. NAC delivers endpoint protection, access control, and performance monitoring, authentication, and network security enforcement as shown in Fig. 1. As shown in the Figure 1, the NAC solution consists of three major components. The first component is a policy decision point (aka Radius Server) which acts as a policy repository and authenticator using Authentication/Authorization/Accounting model (AAA). Secondly, the policy enforcement point which is a network switch that communicates with a radius server to manage the accessibility to the network's resources.

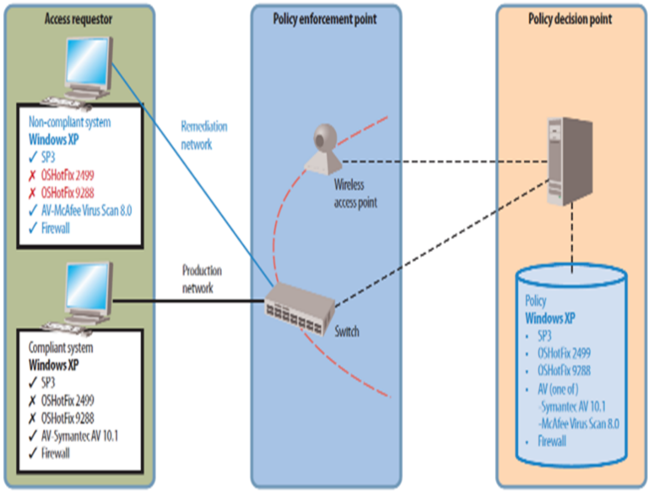


Figure 1: NAC System Components.

Lastly, an endpoint agent is implemented on endpoint devices to check the needed requirements based on pre-defined policies. Network access control compels all network users to adapt to its automated directives aimed to protect the network from security threats. With NAC system deployed, enforcing any anti-malware software on endpoint nodes is a compliment. Hence, the network will automatically take over the task of preventing malware attacks. Data theft and the desire to cause disruption are among the reasons why some attackers raid network access control. Examples of attacks purposed for stealing data are Denial of Service (DoS) and Address

[*] Department of Information Science ,Kuwait University, Kuwait. Email: :kalimuddinqureshi@gmail.com

[†] Department of Information Science ,Kuwait University, Kuwait. Email: :kalimuddinqureshi@gmail.com

[‡] Former-Dean College of Computing Science and Engineering, Kuwait University, Kuwait. Email: : mohsen.alshamali@ku.edu.kw

Resolution Protocol (ARP) spoofing. If a good defiance strategy is deployed, it will make the network impermeable to many of the arms brought about by these attacks [10], [6]. In this paper, an enhanced version of PacketFence is introduced by reconfiguring the policy enforcement point (Cisco Switch) with the best practice guidelines provided by Cisco [4] to reinforce the NAC solution. A certain type of attack has been deployed over the simulated network using GNS3 and Yersinia attacking tool [10]. In this work, CDP attacks and root-claim attacks are introduced which show a vulnerability in a major part of the network access control system (unmodified PacketFence), also known as an enforcer device.

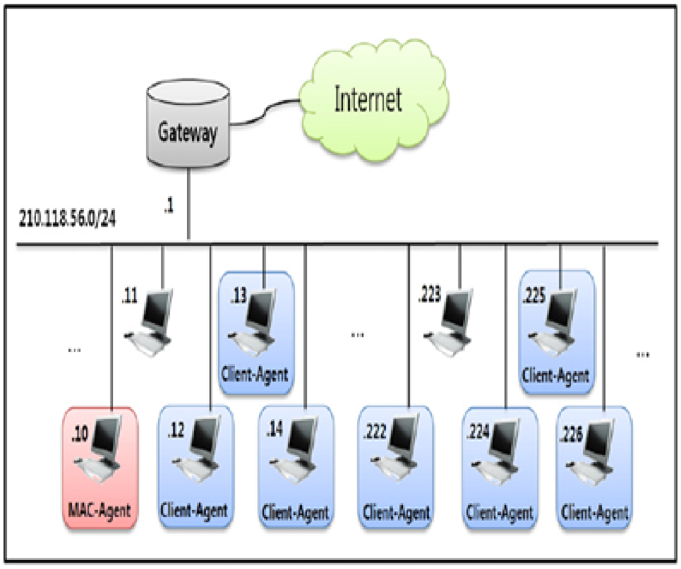


Figure 2: ARP spoofing suggested prevention method.

The paper is organized as follows: the literature review is presented in Section 2. In Section 3, NAC Tools, and an enhanced version of PacketFence is presented by configuring the policy enforcement point along with the implementation. The core setup and the tools used tobuild a network simulation are presented in Section 4. Experimentation and evaluation results are shown in Section 5, along with the performance metrics used. Finally, a closure of the research and future works is presented in Section 6.

2 Literature review

The study in [5] presents a method to lower the rate of attacks based on ARP spoofing. Along with that, a way to include stronger security measures for basic control systems without any additional cost is discussed. The method requires no changes of the required protocol or any extra appliances to prevent ARP spoofing. It only requires an actual detached PC with a MAC-Agent and a Client-Agent in each PC in the network, as shown in Fig. 2. The study in [7] points out consideration of the standards of conduct of the organization and its clients. It conveys an improved organization access control utilizing free BSDpfSense open-source systems. It uses the a committed edge firewall with the presentation of squid, squidGuard, Squid Analysis Report Generator (SARG), as well as the establishment of an Active Directory worker with client access arrangements to improve client access control and protect the LAN from abuse, virus attacks, and unauthorized entries. As a result of this study, two main problems were observed. The first is the limited bandwidth, and the second is the absence of a reliable network access control, which left the UMaT network exposed to all types of attacks such as DoS/DDoS attacks, Worms, Trojans attacks, etc. By applying a proxy server (Squid) which limits access to the network, the UMaT network has been improved, and network vulnerability has been greatly reduced.

The study in [3] contains the evaluation of the security provided by the PacketFence Network Access Control server installed within the GNS3 emulator. To prove that the attacks on the network are real, open-source software called Yersinia has been used to conduct various experiments. As a result, the study showed that the PacketFence system is vulnerable, and the attacks are viable and realistic, which may lead to a considerable cost during the attacks for the company. Additionally, the study discussed various types of attacks on the network; for instance, CDP attack, MAC Flooding attack, Authentication attack, and ways to secure the network against them.

Open NAC solutions are hybrid solutions composed of software and hardware components. The most popular open-source NAC solutions are cited in [10]. A number of researchers use open-source tools for educational purposes due to the availability of code to community. OpenNAC(Opennac.org, 2021) is an open-source network access control system that offers access to LAN/WAN networks based on privilege rights and policies. OpenNAC also proposes a secure connection among devices on the network by using 802.1X authentication protocols based on LDAP or an active directory. Figure 3 provides several features of OpenNAC to manage the network. FreeNAC [10] is a GPL (General Public License) open-source network access control system. It is completely free and supports both wired/wireless network infrastructure. It serves all different types of network devices while focusing on information security when communicating with different devices on the network. It achieves this by applying security modes such as 802.1X, MAB (MAC-Authentication-Bypass), and VMPS (VLAN management policy server). FreeNAC is hard to perform a posture assessment and bandwidth monitoring. Therefore, it is relying on other tools, for example, security assessment tools, from the server side and bandwidth monitoring tools. Please refer toFig,3 for an architecture of the tool.

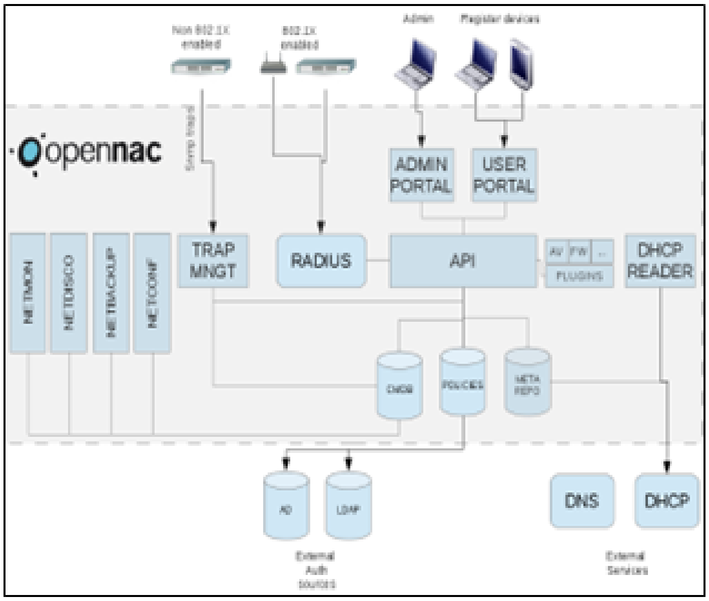


Figure 3: OpenNAC design architecture

PacketFence (packetfence.org, 2021) can be used in both wired/wireless networks with a unified management. It also uses a secure connection between the NAC elements in the network by using 802.1X, MAC-Authentication-Bypass (MAB), DHCP fingerprinting, and user agents which can be installed in the endpoint device. The system also uses a posture assessment which can be defined as a compliance verification to evaluate the endpoint device security perspective based on the pre-defined rules and policies in the PacketFence server. Moreover, it performs posture assessment using statement of health (SOH) protocols to collect the required data from devices in the network. PakcetFence can perform remediation through a captive portal and redirect the user to a different URL with a set of instructions for the specific situation to get access to the network. The basic implementation for deploying PacketFence consists of 3 major components, as shown in Fig.4.

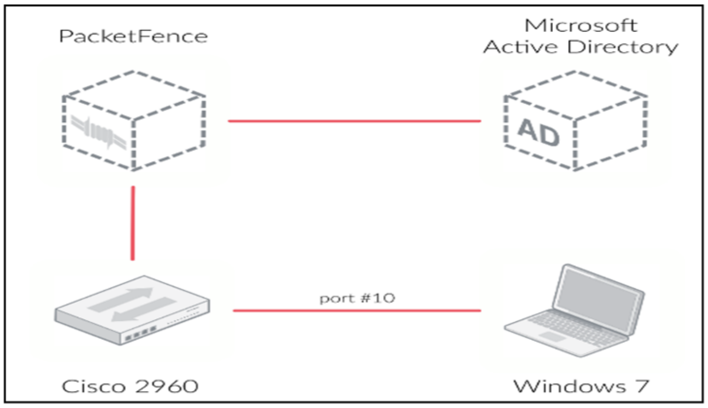


Figure 4: Basic implementation of PacketFence

The architecture design of PacketFence (see Figure 5) can analyse the traffic bandwidth and keeping track of it in case of any unusual or suspicious activity. Additionally, the system performs a set of actions to secure the network by quarantine or changing the access level of the device.

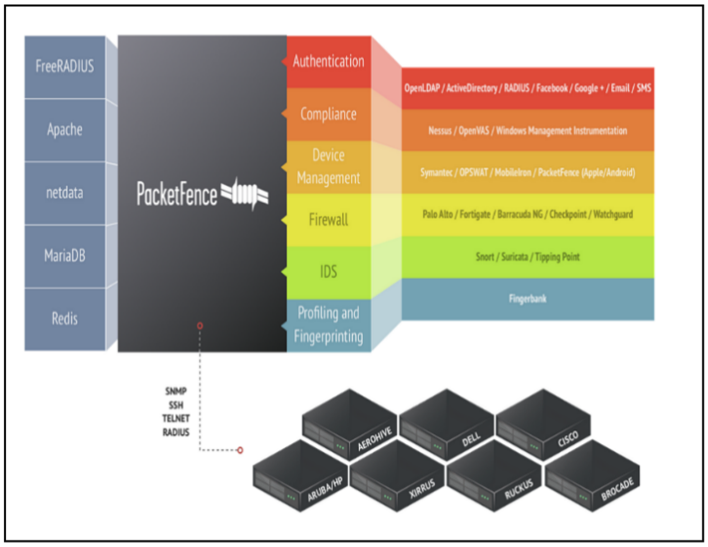

FreeRADIUS
Apache
netdata
MariaDB
Redis
PacketFence
Authentication
Compliance
Device Management
Firewall
IDS
Profiling and Fingerprinting
OpenLDAP / ActiveDirectory / RADIUS / Facebook / Google + / Email / SMS
Nessus / OpenVAS / Windows Management Instrumentation
Symantec / OPSWAT / MobileIron / PacketFence (Apple/Android)
Palo Alto / Fortigate / Barracuda NG / Checkpoint / Watchguard
Snort / Suricata / Tipping Point
Fingerbank
SNMP
SSH
TELNET
RADIUS
AEROHIVE
DELL
CISCO
ARUBA/HP
XIRRUS
RUCKUS
BROCADE

Figure 5: PacketFence - Component Architecture

Table 1 provides a summary comparison made based on a study paper reviewing open-source network access control tools for Enterprise educational networks [10]. The table consists of features that are mostly embedded with the security aspect of NAC tools.

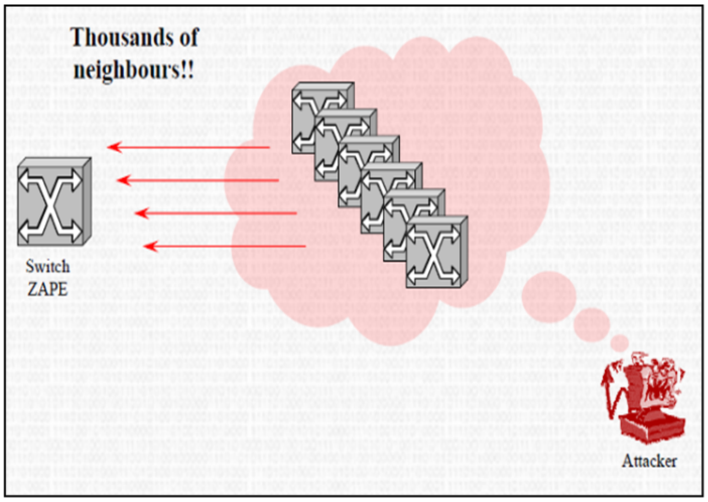


Figure 6: A basic illustration of CDP attack [6, 7]

From the table 1 and 2 it clear that PacketFence can manage the bandwidth, a robust posture analysis, multiple authentication protocols, and standards. While PacketFence succeeds other open-source solutions certain issues need to be tackled. In [3]. PacketFence is vulnerable to certain types of attacks; as a result, an eagle eye is needed to focus on this issue. Starting from

Table 1: A Comparison between open-source NAC security aspects

Feature	OpenNAC	PacketFence	FreeNAC
Tracking Bandwidth	Doesn't keep track of bandwidth usage	Keep track of bandwidth usage and takes action on any suspicious activity such as quarantine.	Relying on other Network monitoring tools to keep track of bandwidth usage
Authentication	802.1X based on LDAP protocol	MAB "MAC-Authentication bypass" 802.1X DHCP fingerprinting User agents	MAB "MAC-Authentication bypass" 802.1X, Cisco VMPS (Vlan management policy)
Posture assessment	Examine end devices' antivirus updates, OS updates, patches, and firewall	Use SOH protocol "Statement of health" to gather data from end-devices and analyze the posture	Unable to do posture assessment by itself, and relying on other security assessment tools from the server side to deploy posture assessment for end-devices
Wire/Wireless Networks	Supported	Supported	Supported
Design Scalability	Small/Medium Networks	Geographically Scalable	Small Networks

the results of this study, we introduce in the next section an enhanced version the PacketFence system.

3 The Proposed Enhancements to the PacketFence System

While PacketFence has the overall success among others, nevertheless, it has been shown that the system can fail due to certain vulnerabilities. Here, improving the security demands focusing on hardening a major component of the PacketFence system can be achieved by providing the configuration and implementation of the policy enforcement point (Network switch) where any attack against the switch can eventually lead to a system failure, fore more detail refer to Section 6.

Regarding the vulnerabilities on policy enforcement points, the following section will analyse the introduced attacks and prevention methods according to the best practice guidelines, as provided by the Cisco official website [4]. In the upcoming sub-sections, a set of attacks is exposed beside the resolving configuration for securing the policy enforcement point.

Table 2: A comparison of features between open-source NAC solutions

Features	OpenNAC	PacketFence	FreeNAC
Posture Analysis	Fairly	Yes	Not Inherent
Contrivance Authentication	Yes	Yes	Yes
Bandwidth Management	No	Yes	No
Network Vendor Support	Multivendor	Multivendor	Multivendor
Wired and Wireless Support	Yes	Yes	Yes
Software Integration	Yes	Yes	Commercial
Community Support	Active	Active	Fairly Active
Administrative Interface	Web Interface	Web Interface	Mainly Window-based
Reporting	Yes	Yes	Yes

3.1 CDP Attack

CDP (Cisco Discovery protocol) is a layer 2 protocol used in cisco devices which sends identification packets over the network in plain text. This protocol is usually used by network administrators for discovering neighbor devices and troubleshooting. The protocol is enabled by default in network devices such as routers, switches, and servers.

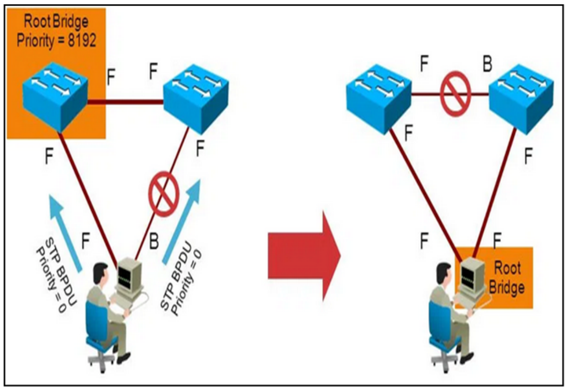


Figure 7: A basic illustration of STP attack (common attack types on switches, 2018)

3.2 STP Attack

Spanning Tree Protocol (STP) is a Layer 2 protocol that operates on switches. Primarily, STP is used to guarantee that you do not make loops when you have duplicate paths in your network; as a result, loops are fatal to the network stability.

Spanning Tree does not use multiple links to lead to the same destination. In addition, Spanning Trees is used in a network fault tolerance design in such way if one link dropped another backup link will take place to improve reliability and resilience. STP attacks (figure 7) focus on spoofing the root bridge in the network topology by broadcasting out a topology change enforcing an STP re-calculation. This also causes a DoS on the network by causing an interruption during the root bridge modifications, as shown in the below fig. 7.

In this kind of attack, mitigation can be attained by not using STP on unnecessary ports disable stp and using port security commands on the interface level. Moreover, the bridge protocol data unit (BPDU) guard bpdu guard command is used if the network is using a portfast feature in STP configuration. A command line root guard is mandatory to prevent the attacker to claim the root.

3.3 DHCP Starvation Attack

Dynamic Hosting Configuration Protocol (DHCP) is a layer 2 network management protocol based on a client/server model. The main purpose of DHCP is dynamically assigning an IP address to network devices by a set of requests between the network device and the server. As shown in Fig. 8 DHCP can be implemented on any network, regardless of the size of the network. It can work on layer 3 IP protocols when routers or gateways act like DHCP servers to receive globally unique IP addresses. DHCP Starvation Attack is a malicious attack often used to exhaust the DHCP server by sending numerous requests. This attack aims to stall the network by preventing legitimate devices from acquiring an IP address to gain access to the network. In this scenario, a prevention method is introduced by applying an ip dhcp snooping command on the switch, which drops undesirable DHCP traffic that may be requested from unauthorized DHCP servers [13].

3.4 DTP Attack

The Dynamic Trunking Protocol (DTP) is a layer 2 cisco proprietary protocol that can only be used between two cisco switches to negotiate a trunk link between them along with the encapsulation type. DTP is automatically enabled on a switch port by default when a certain trunking mode is configured on the switch port. There are two DTP trunking modes: first is dynamic desirable, and the other is Dynamic auto. For the first mode, the interface sends DTP packets constantly trying to establish the trunk link if possible. Meanwhile, in the dynamic auto mode, the interface just waits for DTP requests that are being sent. An attack can be launched to force interfaces that use DTP protocol to shift to trunk mode. This will leave the network vulnerable to various types of attacks such as traffic sniffing, man in the middle attacks, and VLAN hopping as a result of all VLANS being accessible for the attacker.

3.5 VTP Attack

VLAN Trunking Protocol (VTP) is another Cisco proprietary layer 2 protocol based on the client/server model which is used to broadcast VLAN information such as VLAN name and ID across the network. It has been introduced to synchronize VLAN information with all switches inside the network by using the same VTP domain and password. This protocol helps network administrators to efficiently manage VLANs with less time by creating or deleting a VLAN in one switch and syncing this information across all the switches in the network instead of creating or deleting VLANs in every switch.

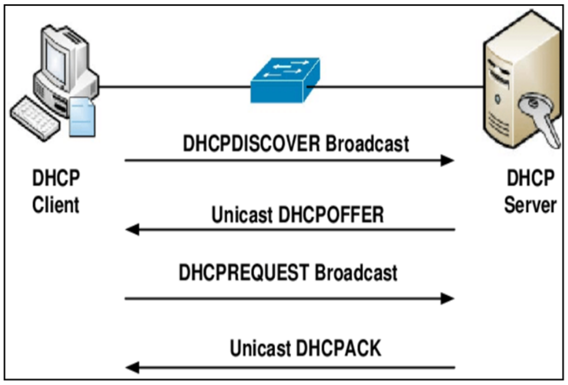


Figure 8: Dynamic Hosting Configuration Protocol is working between the client and the server [8, 9]

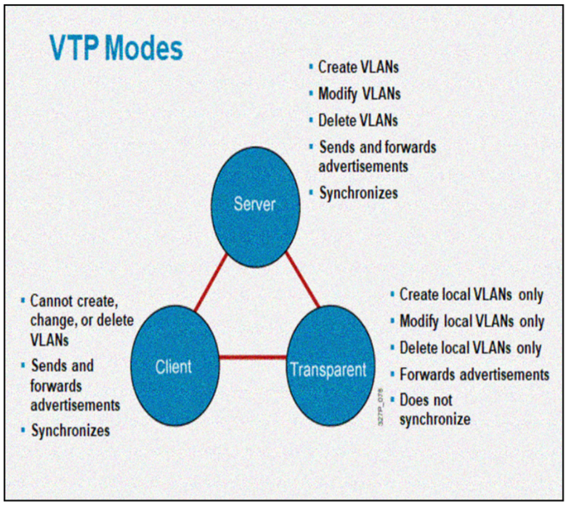


Figure 9: Virtual Trunking Protocol Modes [10]

As shown in figure 9, there are three VTP modes which can be configured in Cisco switches:

1- VTP server: Usually configured in the core switch that will advertise VLAN information across the network.

2- VTP client: Enable the switch to collect and save VLAN information.

3- VTP Transparent: The switch will be able to resend the VLAN information from the VTP server and, at the same time, can create or delete VLANs locally in it. However, the switch will not sync on its own VLAN information to the network.

VTP attack methods launch on the network through modifying the switch's VLAN, either by adding a new VLAN or deleting an existing VLAN. This could cause a disruption in the whole network. Therefore, to protect the switch, the network administrator should use a VTP password and disable VTP protocol if not required.

4 Experimental Setup

The need for a hosting machine is essential for preparing the development environment to deploy the enhanced proposed solution (enhanced PacketFence). The core setup consists of three parts. Firstly, the hosting machine and its specifications. Secondly, software tools such as operating systems, Yersinia security attack tool, and the open-source NAC solution. Finally, a GNS3 network simulator. The reader should notice that the core switch (policy enforcement point) is integrated with a GNS3 network simulator. We explain this in the following sections.

4.1 Hosting configuration

In this part a set of tools and machine configurations are introduced that describe how the environment is prepared and settled is explain below.

1. Hosting Machine The hosting machine is a PC with Windows Server 2016 R2 64-bit OS. The processor is Intel Xeon E5-2687W v2 (dual processor) running @ 3.40 GHz. The installed RAM is 128 GB DDR3. It also has an Intel I210 Gigabit network connection ethernet adapter for network communication.

2. Operating System and Software tools PacketFence is the network access control solution used in the development environment along with Yersinia (Network Attacking tool). PacketFence can be downloaded from the official website (packetfence.org, 2021) as a standalone image or bundled with pre-deployed Linux OS images. In the case of using the PacketFence bundled image, a virtual machine is needed such as oracle virtual box is needed to import the image itself. In contrast, the standalone image needs to be installed manually in Linux OS. Both flavours need Linux OS. Yersinia [12] is an attacking open-source tool used to produce various types of attacks including CDP, root claim, DHCP starvation, floods, and so many others. Yersinia uses libpcap, ncurses, GUI, and libnet dependencies to function properly.

3. Network Simulator tool (GNS3) GNS3 is an open-source network simulator tool written in Python and can be downloaded from their official website (gns3.org, 2021). GNS3 can be integrated with many virtual machines including network components such as switches, routers, DHCP servers, firewalls, radius servers, and so many others. In addition, a variety of VM images, including operating systems, application servers, PacketFence, network firmware, and more, could be imported in GNS3. An Application Programming Interface (API), the so-called libpcap, is used to communicate between network components and endpoints. The product version 2.2.15 is used to deploy the environment.

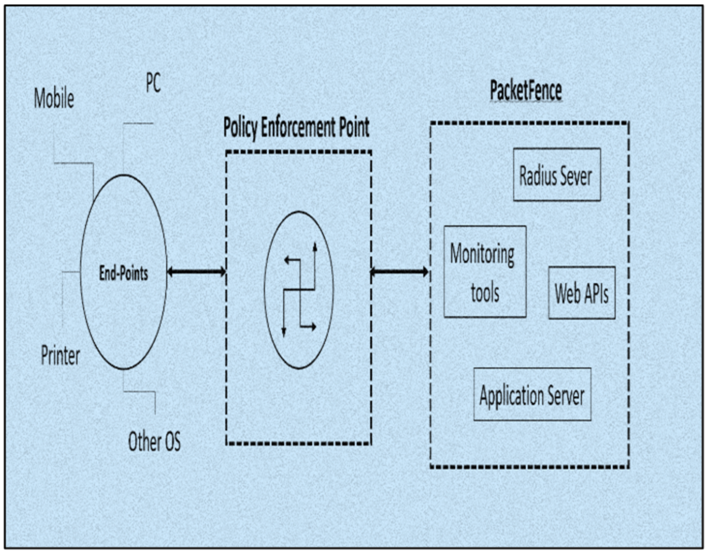


Figure 10: The relationship of policy enforcement points between PacketFence and endpoints

4.2 System Deployment

An overall deployment of network components, attacks, and configurations is presented in the following manner: network topology implementation, policy enforcement point configuration, launch of network attacks, and finally, hardening configuration on the policy enforcement point. In Fig. 10 a diagram is presented to illustrate the PacketFence system and how it communicates with the endpoint devices. PacketFence communicates with the policy enforcement point via a radius server as a built-in component to apply the pre-defined policies stored in the application server on the endpoint devices. It has its own application server and web APIs to configure and fetch information from devices in the network, such as the device's MAC address, IP, operating system, etc. This kind of information helps the PacketFence system set a profile for each endpoint device in the network and design a specific role for permitting/denying access to the network. Moreover, the system has an embedded monitoring tool which can perform on different levels: for example, at the system level, such as CPU utilization, Disk space level, system's RAM, and more. On the radius and authentication level, it shows the radius server latency, the number of requests on the radius, and the number of successful and unsuccessful authentications that occurred.

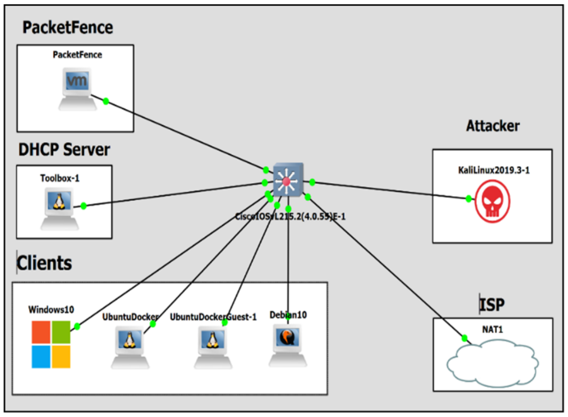


Figure 11: A network topology implemented in GNS3 simulator

The system can monitor the endpoint devices using SOH (Statement of health) protocols. It shows the number of devices on the network, security events that may happen when unregistered devices try to access the network, changes that take place on an authorized device (i.e., changing the device OS), or if the antivirus software is not up to date. In addition, the system can delegate unauthorized devices to a posture assessment process by redirecting the device to a defined URL which provides a list of guidelines to accommodate the network.

4.3 Network Topology Implementation

Network topology is an essential arrangement structure of network elements that aids network administrators to analyse the data flow, physical/logical interconnections, and transmission rates of the network. The network environment implementation forms a star topology. Mainly, the environment is composed of six elementary entities, as shown in Fig. 11.

(a) The open source NAC software solution (PacketFence). The DHCP server responsible for assigning IP addresses to network devices with a range of 192.168.122.0/24.

(b) The attacker is represented by Kali OS, which has great frameworks to launch a variety of different types of attacks. In this thesis, the Yersinia framework is used to launch various network attacks, as will be shown in part 3 of this section.

(c) GNS3 uses a NAT (Network Address Translation) to supply the internet connection to the network.

(d) Several nodes (clients) are used with different operating systems acting as endpoint devices.

(e) Finally, one of the major parts of the NAC solution (policy enforcement point) is represented by a layer 2 Cisco switch which applies the pre-defined policies inherited from PacketFence software on clients.

4.4 Policy Enforcement Point Configuration and Connectivity

In Fig. 12 we show the communication between PacketFence and the policy enforcement point.

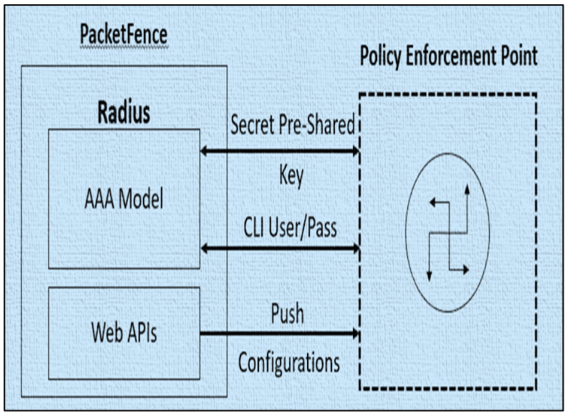
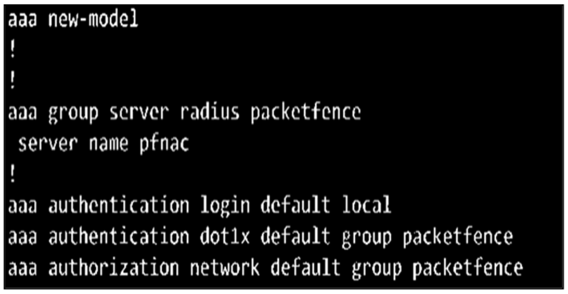


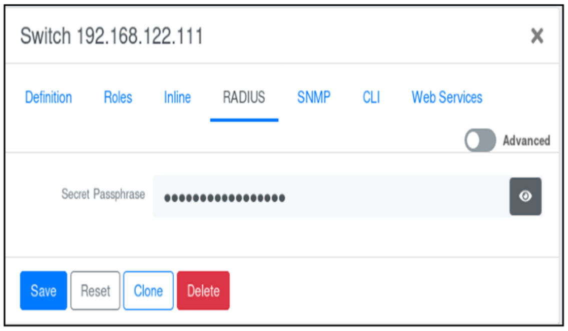
Figure 12: A communication between PacketFence and Policy Enforcement Point

The PacketFence initiates communication link between the embedded radius server and the policy enforcement point using AAA model (Figure 13).



```
aaa new-model
!
!
aaa group server radius packetfence
 server name pfnac
!
aaa authentication login default local
aaa authentication dot1x default group packetfence
aaa authorization network default group packetfence
```

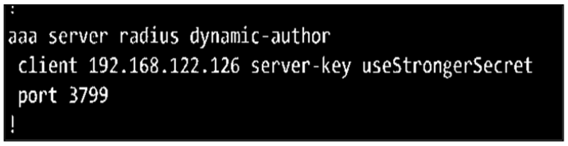
Figure 13: AAA model configuration in policy enforcement point



Switch 192.168.122.111 ×

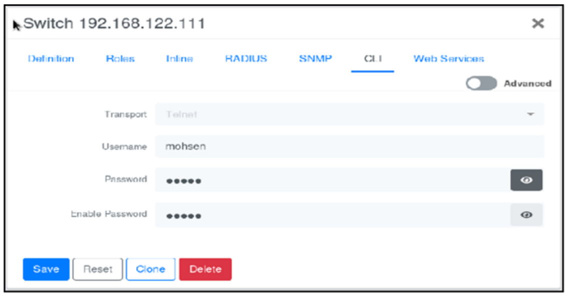
Secret Passphrase ●●●●●●●●●●●●●●●●

(a)PacketFence UI



```
!
aaa server radius dynamic-author
 client 192.168.122.126 server-key useStrongerSecret
 port 3799
!
```

(b) The policy enforcement points of applying pre-shared key



Switch 192.168.122.111

Transport: Telnet

Username: mohsen

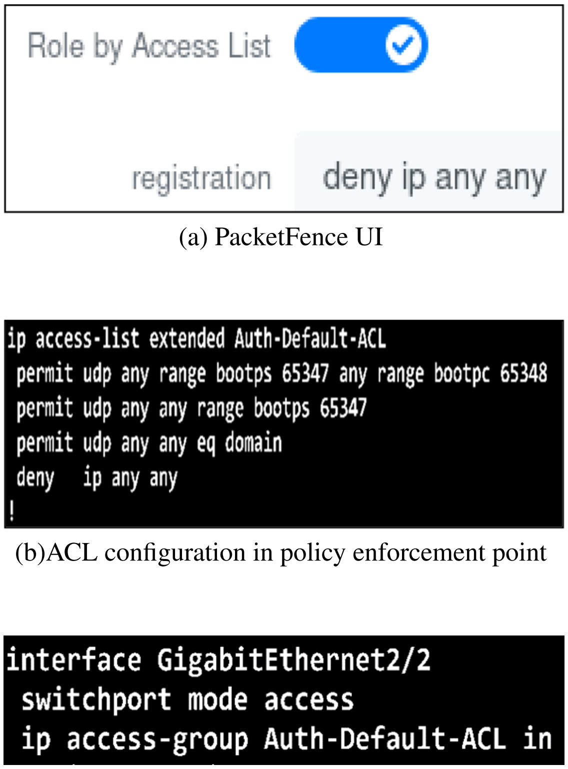
Password: •••••

Enable Password: •••••

(c) CLI Authentication configuration in PacketFence UI

Figure 14: Radius configuration in PacketFence UI

This can be done by setting a pre-shared key that should be configured in both PacketFence and the policy enforcement point as well as the CLI's (Command Line Interface) username and password. This way, the system can push the configuration to the policy enforcement point using the Web APIs, as shown in Figure 14. One of the methods used to control access to the network is the use of ACL (Access Control List). ACL is one of the essential structures in network configuration using permit/deny rules to manage network accessibility (Cisco CCNA/CCENT Exam 640-802, 640-822, 640-816 Preparation Kit, 2009). PacketFence uses ACL which can be defined in the system's UI and then pushed to the policy enforcement point (see figure 4.3) with the pre-defined rules as shown in Figure 15.



Role by Access List

registration: deny ip any any

(a) PacketFence UI

```
ip access-list extended Auth-Default-ACL
 permit udp any range bootps 65347 any range bootpc 65348
 permit udp any any range bootps 65347
 permit udp any any eq domain
 deny   ip any any
!
```

(b)ACL configuration in policy enforcement point

```
interface GigabitEthernet2/2
 switchport mode access
 ip access-group Auth-Default-ACL in
```

(c) Applying ACL on the switch interface level

4.5 Launch of Network Attacks

In the second part of the previous section, five attacks on a simulated network created on GNS3 had been introduced and discussed. As shown in figure 4, the attacker is assumed to be an end-user with respect to the network. In our scenario, the attacker performs all layer-2 attacks within the network. Technically, these attacks are launched via the Yersinia tool, which is hosted on a virtual machine (Kali Linux distribution). In addition, all network components,including VM machines, are centralized and monitored on GNS3 via a set of APIs and services. All five attacks (CDP – DTP – VTP – STP – DHCP starvation) are directed toward the policy enforcement point (Cisco Switch) where the first four attacks directly disturb the switch. Meanwhile, the remaining one (DHCP starvation) affects the endpoint devices via the switch.

4.6 Hardening Configuration

The main objective in this paper is to present a methodology for hardening an open-source network access control system (PacketFence). Configurations that are used to harden the system will be categorized as follows: configurations recommended by the Cisco official guideline, and a proposed configuration that resolves some of the introduced attacks that Cisco did not include in their guideline [7].

A. Configurations Recommended by Cisco Guideline Here, a resolution will be highlighted for each attack by implementing configurations on the Cisco switch from the Cisco guidelines [7]
.

1. CDP Attack

As mentioned in [7] for hardening Cisco devices, CDP protocol should be only enabled on switches that relate to a trusted network. Configurations that need to be applied are.

(a) no cdp enable command in the interface level or using.

(b) no cdp run command on the global level of the switch.

2. DHCP Attack

Cisco guidelines emphasize dropping undesirable DHCP traffic requested from unauthorized DHCP servers by executing ip dhcp snooping commands on the switch.

B. Proposed Configurations

1. STP Attack

Mitigation of STP attacks can be accomplished by disabling STP on unnecessary ports using disable stp command and port security command on the interface level. In case the network is using a port-fast feature in STP configuration, a bridge protocol data unit (BPDU) guard bpdu guard command is used. Moreover, root guard commands are applied to prevent the attacker from claiming the root.

2. DTP Attack A resolution to this attack is achieved by turning off DTP protocols on all switch ports by executing the switchport nonnegotiate switch command. This turns all ports from auto negotiations to off.

3. VTP Attack

In this attack, a precaution is needed for protecting the network VLANs from being manipulated (add/delete/modify).

It is required from the administrator to enable a VTP password command and to stop using VTP protocols if not needed. Furthermore, a good practice to follow is to use VTP modes wisely based on the criteria of the environment.

Table 3: CDP attack affects CPU Utilization over time

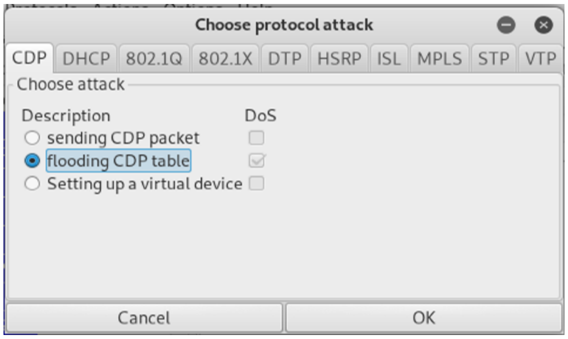
Minute	Number of packets	CPU Utilization
1	28,000	22%
3	85,000	31%
5	150,000	34%
7	210,000	72%
10	375,000	84%
Average		48.6%

5 Obtained Results and Discussion

In this part of the paper, we will show the results obtained in the research before and after introducing several network attacks. The dataset of the research is made up of five attacks tackling a major part of the PacketFence solution (the Policy Enforcement Point) which is a Cisco network switch under model numbers 2960, 2970, 3560 and 3750.

Yersinia is a leading tool to generate attacks and is heavily used in security network research (Opennac.org, 2021). The UI of Yersinia follows a WYSIWYG [[12], [8], [11]]. See Figure 17. All attacks introduced here are generated using the Yersinia tool operated on Kali Linux OS. Next, objective, and subjective results will be shown for each attack.

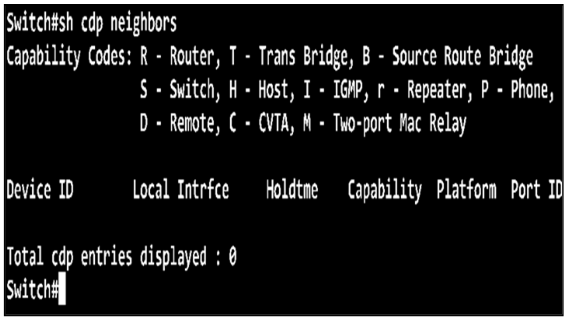
5.1 Cisco Discovery Protocol Attack Evaluation As we mentioned earlier, a CDP attack is launched using the Yersinia tool attacking the policy enforcement point. Figure 18 shows an empty table as an initial condition of neighbor devices that use CDP protocol. By contrast, figure 19 shows the generated devices from launching a CDP flood attack.



Choose attack

Description	DoS
○ sending CDP packet	□
● flooding CDP table	☑
○ Setting up a virtual device	□

fig. 16.Yersinia UI for launching various type of attacks.



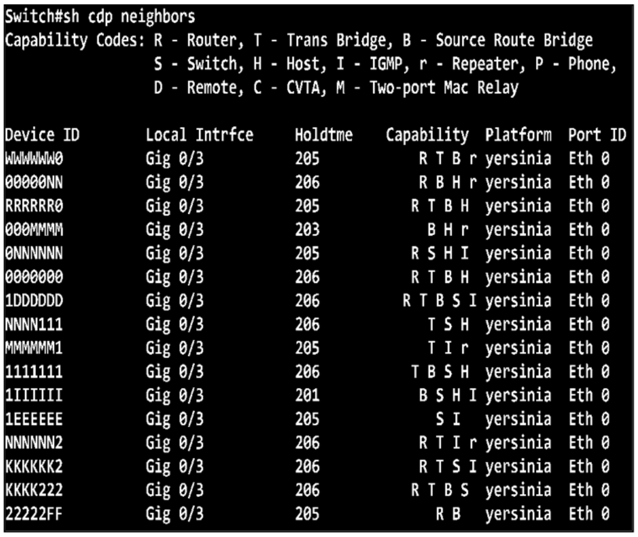
```
Switch#sh cdp neighbors
Capability Codes: R - Router, T - Trans Bridge, B - Source Route Bridge
                  S - Switch, H - Host, I - IGMP, r - Repeater, P - Phone,
                  D - Remote, C - CVTA, M - Two-port Mac Relay

Device ID       Local Intrfce     Holdtme    Capability  Platform  Port ID

Total cdp entries displayed : 0
Switch#
```

Fig. 17. A switch shows a CDP neighbor before launching the attack

The CDP flood attack dramatically increases the CPU utilization of the switch as the number of packets increases over time, as shown above in table 3, and is illustrated in the chart of figure 20 (e). The CDP attack is launched over time at different durations, starting from 1 minute up to 10 minutes. Figure 20 (a) shows the CPU utilization before starting the attack while (d) shows the CPU utilization after the end of the attack. As shown in (e), the attack significantly decreases the performance of the switch as CPU utilization increases for a period.

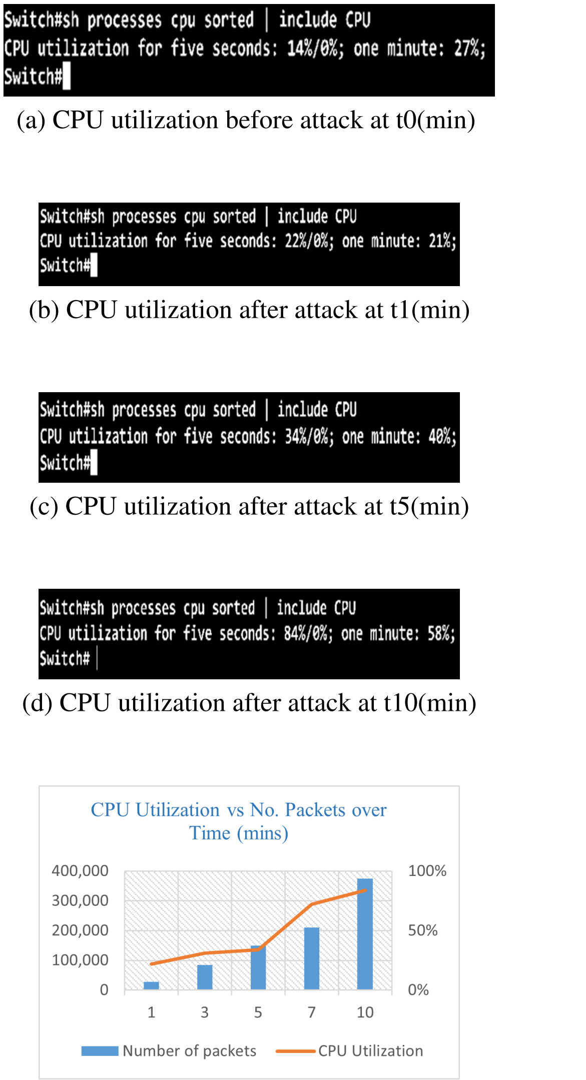
5.2 Dynamic Hosting Configuration Protocol Attack Evaluation

Another attack that can be generated using Yersinia is a DHCP starvation attack which can be avoided by applying a hardening configuration. The following figure 21 shows how the attack prevents the endpoint devices from obtaining IP addresses to access the network. After launching the attack, the dedicated process in the end-device system, which is responsible for obtaining IP addresses, is stalled due to the exhausted DHCP server, which gets numerous requests.



```
Switch#sh cdp neighbors
Capability Codes: R - Router, T - Trans Bridge, B - Source Route Bridge
                  S - Switch, H - Host, I - IGMP, r - Repeater, P - Phone,
                  D - Remote, C - CVTA, M - Two-port Mac Relay

Device ID        Local Intrfce     Holdtme    Capability  Platform  Port ID
WWWWWW0          Gig 0/3           205          R T B r yersinia  Eth 0
00000NN          Gig 0/3           206          R B H r yersinia  Eth 0
RRRRRR0          Gig 0/3           205         R T B H  yersinia  Eth 0
000MMMM          Gig 0/3           203           B H r  yersinia  Eth 0
0NNNNNN          Gig 0/3           205         R S H I  yersinia  Eth 0
0000000          Gig 0/3           206         R T B H  yersinia  Eth 0
1DDDDDD          Gig 0/3           206        R T B S I yersinia  Eth 0
NNNN111          Gig 0/3           206           T S H  yersinia  Eth 0
MMMMMM1          Gig 0/3           205           T I r  yersinia  Eth 0
1111111          Gig 0/3           206         T B S H  yersinia  Eth 0
1IIIIII          Gig 0/3           201          B S H I yersinia  Eth 0
1EEEEEE          Gig 0/3           205            S I   yersinia  Eth 0
NNNNNN2          Gig 0/3           206          R T I r yersinia  Eth 0
KKKKKK2          Gig 0/3           206          R T S I yersinia  Eth 0
KKKK222          Gig 0/3           206         R T B S  yersinia  Eth 0
22222FF          Gig 0/3           205            R B   yersinia  Eth 0
```

Fig. 18. A switch shows a CDP neighbor after launching the attack with dummy generated devices



```
Switch#sh processes cpu sorted | include CPU
CPU utilization for five seconds: 14%/0%; one minute: 27%;
Switch#
```

(a) CPU utilization before attack at t0(min)

```
Switch#sh processes cpu sorted | include CPU
CPU utilization for five seconds: 22%/0%; one minute: 21%;
Switch#
```

(b) CPU utilization after attack at t1(min)

```
Switch#sh processes cpu sorted | include CPU
CPU utilization for five seconds: 34%/0%; one minute: 40%;
Switch#
```

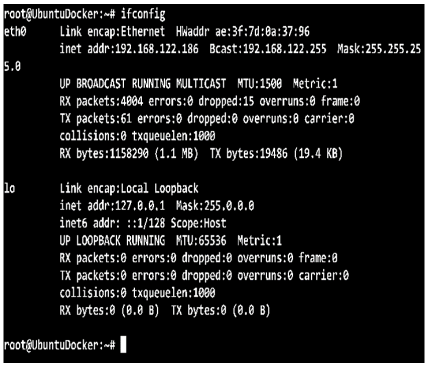
(c) CPU utilization after attack at t5(min)

```
Switch#sh processes cpu sorted | include CPU
CPU utilization for five seconds: 84%/0%; one minute: 58%;
Switch#
```

(d) CPU utilization after attack at t10(min)

(e) A chart illustrates a CDP attack affects the CPU utlilization over time

Fig. 19. Switch shows a CPU utilization before and after launching attack

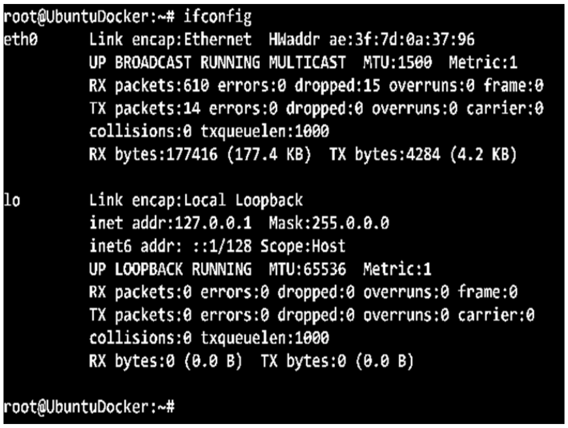


```
root@UbuntuDocker:~# ifconfig
eth0      Link encap:Ethernet  HWaddr ae:3f:7d:0a:37:96
          inet addr:192.168.122.186  Bcast:192.168.122.255  Mask:255.255.25
5.0
          UP BROADCAST RUNNING MULTICAST  MTU:1500  Metric:1
          RX packets:4004 errors:0 dropped:15 overruns:0 frame:0
          TX packets:61 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:1158290 (1.1 MB)  TX bytes:19486 (19.4 KB)

lo        Link encap:Local Loopback
          inet addr:127.0.0.1  Mask:255.0.0.0
          inet6 addr: ::1/128 Scope:Host
          UP LOOPBACK RUNNING  MTU:65536  Metric:1
          RX packets:0 errors:0 dropped:0 overruns:0 frame:0
          TX packets:0 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:0 (0.0 B)  TX bytes:0 (0.0 B)

root@UbuntuDocker:~# 
```

(a) An endpoint device already obtained IP address before the attack

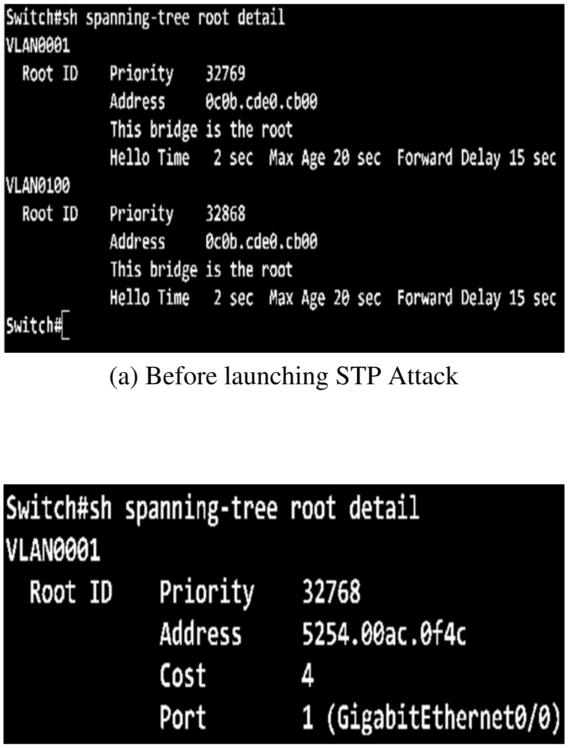
5.3 Spanning Tree Protocol Attack Evaluation

As discussed previously, this attack works on changing the cost of the switch and trying to alter the topology of the network. In figure 22, the initial cost will be zero. As attacks are involved, the cost will be changed as shown in (a) and (b). Moreover, the attack changes the traffic bandwidth interface port. As the number of attacks passes, the root cost of the device (switch) is accordingly changed, as shown in table 4. The assumption here is that the attack is launched with a secondary edge switch connected to the policy enforcement point switch (the device).



```
root@UbuntuDocker:~# ifconfig
eth0      Link encap:Ethernet  HWaddr ae:3f:7d:0a:37:96
          UP BROADCAST RUNNING MULTICAST  MTU:1500  Metric:1
          RX packets:610 errors:0 dropped:15 overruns:0 frame:0
          TX packets:14 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:177416 (177.4 KB)  TX bytes:4284 (4.2 KB)

lo        Link encap:Local Loopback
          inet addr:127.0.0.1  Mask:255.0.0.0
          inet6 addr: ::1/128 Scope:Host
          UP LOOPBACK RUNNING  MTU:65536  Metric:1
          RX packets:0 errors:0 dropped:0 overruns:0 frame:0
          TX packets:0 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:0 (0.0 B)  TX bytes:0 (0.0 B)

root@UbuntuDocker:~#
```

(b) Same endpoint device is stalling and unable to obtain IP address due to DHCP attack Figure 20. A DHCP attack is applied and stalls the process of obtaining IP address in the system

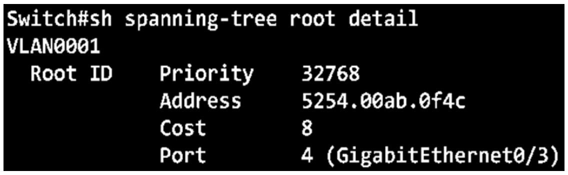


```
Switch#sh spanning-tree root detail
VLAN0001
  Root ID    Priority    32769
             Address     0c0b.cde0.cb00
             This bridge is the root
             Hello Time   2 sec  Max Age 20 sec  Forward Delay 15 sec
VLAN0100
  Root ID    Priority    32868
             Address     0c0b.cde0.cb00
             This bridge is the root
             Hello Time   2 sec  Max Age 20 sec  Forward Delay 15 sec
Switch#
```

(a) Before launching STP Attack

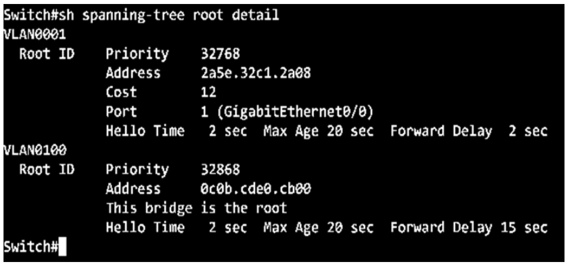
```
Switch#sh spanning-tree root detail
VLAN0001
  Root ID    Priority    32768
             Address     5254.00ac.0f4c
             Cost        4
             Port        1 (GigabitEthernet0/0)
```

(b) 1st occurrence of STP attack by Yersinia



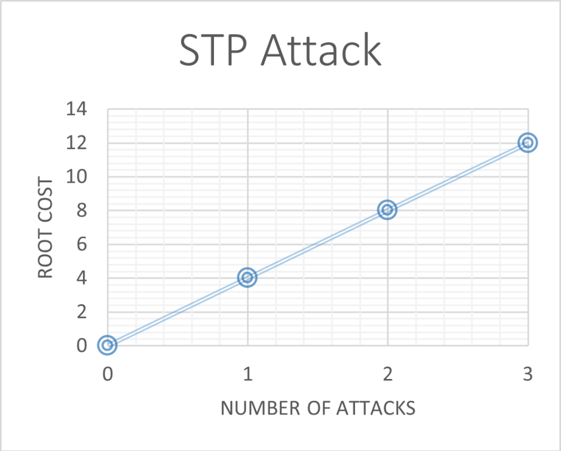
```
Switch#sh spanning-tree root detail
VLAN0001
  Root ID    Priority    32768
             Address     5254.00ab.0f4c
             Cost        8
             Port        4 (GigabitEthernet0/3)
```

(c) 2nd occurrence of STP attack.



```
Switch#sh spanning-tree root detail
VLAN0001
  Root ID    Priority    32768
             Address     2a5e.32c1.2a08
             Cost        12
             Port        1 (GigabitEthernet0/0)
             Hello Time   2 sec  Max Age 20 sec  Forward Delay  2 sec
VLAN0100
  Root ID    Priority    32868
             Address     0c0b.cde0.cb00
             This bridge is the root
             Hello Time   2 sec  Max Age 20 sec  Forward Delay 15 sec
Switch#
```

(d) 3rd occurrence of STP attack.



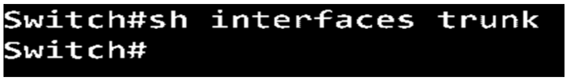
(e) A chart illustrated the occurrence of attacks and the root cost.

Fig. 21. The root cost of the device is affected due to an STP attack.

Table 4: Number of Attacks compared with changing root cost after launching an STP Attack

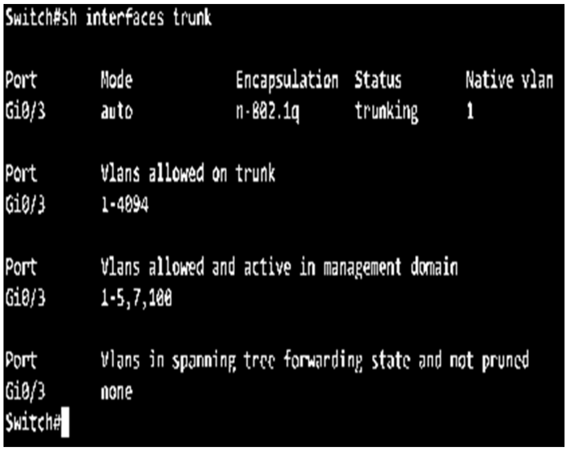
STP Attack Number	Root Cost
0	This bridge is the root
1	4
2	8
3	12
Average	8

5.4 Dynamic Trunking Protocol Attack Evaluation The idea of this attack is to force interfaces that use DTP protocols to be operated on trunk mode. This will leave the network vulnerable to various attacks. Figure 22 shows the switch before and after the attack.



```
Switch#sh interfaces trunk
Switch#
```

(a) Before launching DTP attack where no interface status is trunk.



```
Switch#sh interfaces trunk

Port      Mode         Encapsulation  Status     Native vlan
Gi0/3     auto         n-802.1q       trunking   1

Port      Vlans allowed on trunk
Gi0/3     1-4094

Port      Vlans allowed and active in management domain
Gi0/3     1-5,7,100

Port      Vlans in spanning tree forwarding state and not pruned
Gi0/3     none
Switch#
```

(b) The switch has been forced to enable Trunking mode under port Gi0/3 after launching attack.

Fig. 22. A DTP attack is applied, forcing the switch to enable trunk mode.

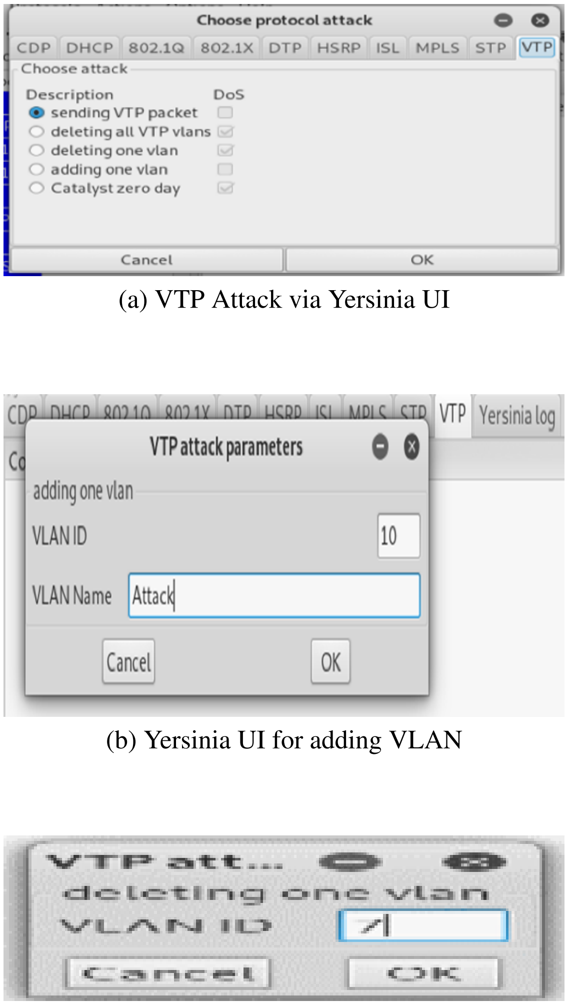
5.5 VLAN Trunking Protocol Attack Evaluation

A VTP attack alters the switch's VLAN, either by adding or deleting the VLAN entry. Some advanced methods can modify existing VLAN entries. This will lead to a disruption in the network. Figure 23 explains the status of the switch VLAN entries before and after launching the VTP attack. As shown in the figure, (a) No added/deleted entries or (b) Adding VLAN 10 under the name Attack to ensure the attack or (c) Deleting an entry in the VLAN table such as VLAN with id 7.

6 Conclusions

Although PacketFence is user-friendly and easy to integrate with many network components such as switches, proxies, and application servers such as radius and DHCP servers, the open-source solution has a wide array of issues that need to be altered and fixed. The second part of this research has been focused on a major part of open-source NAC solutions which is called the policy enforcement point, and which acts as a core switch for the overall solution. In this part, vulnerabilities were discovered after launching a set of attacks on the policy enforcement point component with the help of the Yersinia tool. The regular implementation of the policy enforcement points in open-source solutions (e.g. PacketFence) is missing a set of security guidelines. Although these guidelines are not dependent on starting up the overall solution, the marginalization of these guidelines during deployment with non-trusted networks might produce vulnerabilities.

Hence, hardening methods are represented as commands to enhance the device configuration. The third and final part of this work concentrates on how to harden the policy enforcement point by protecting the network from exposed vulnerabilities, as mentioned in section 4. Guidelines and configurations are presented in this part of the work to tackle these vulnerabilities. As the results shows in section 4, systems that do not follow the recommended guidelines will suffer from exposed vulnerabilities that harness the network device (a.k.a PEP) in terms of CPU utilization, flooding MAC address table, VLAN entries, and root cost, all of which can lead to system failure



(a) VTP Attack via Yersinia UI

CDP DHCP 802.1Q 802.1X DTP HSRP ISL MPLS STP VTP Yersinia log
VTP attack parameters
adding one vlan

VLAN ID 10

VLAN Name Attack

(b) Yersinia UI for adding VLAN

VLAN ID 7

(c)Yersinia UI for deleting entry

Fig. 23. VTP attack is launched via Yersinia, which affects the switch VLAN entries

7 References

References

[1] Bawany, N.Z., Shamsi, J.A., Salah, K.: Ddos attack detection and mitigation using sdn: methods, practices, and solutions. Arabian Journal for Science and Engineering **42**, 425–441 (2017)

[2] Firoozjaei, M.D., Jeong, J.P., Ko, H., Kim, H.: Security challenges with network functions virtualization. Future Generation Computer Systems **67**, 315–324 (2017)

[3] Flores, J., Ramos, V., Lozada, R., Flores, T.: Analysis of solutions of network access control to improve in and out securities on corporative networks. In: 2017 CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON). pp. 1–5. IEEE (2017)

[4] Hong, S., Oh, M., Lee, S.: Design and implementation of an efficient defense mechanism against arp spoofing attacks using aes and rsa. Mathematical and Computer Modelling **58**(1-2), 254–260 (2013)

[5] Inamdar, M.S., Tekeoglu, A.: Security analysis of open source network access control in virtual networks. In: 2018 32nd International Conference on Advanced Information Networking and Applications Workshops (WAINA). pp. 475–480. IEEE (2018)

[6] Iqbal, S., Sujatha, B.: Secure key management scheme for hierarchical network using combinatorial design. Journal of Information Systems and Telecommunication (JIST) **1**(37), 20 (2022)

[7] Iserovich, H.: Empowering network infrastructure Cybersecurity. Ph.D. thesis, The Interdisciplinary Center, Herzliya (2020)

[8] Kim, E., Kim, K., Lee, S., Jeong, J.P., Kim, H.: A framework for managing user-defined security policies to support network security functions. In: Proceedings of the 12th International Conference on Ubiquitous Information Management and Communication. pp. 1–8 (2018)

[9] Liu, D.: Cisco CCNA/CCENT Exam 640-802, 640-822, 640-816 Preparation Kit. Syngress (2009)

[10] Nunoo-Mensah, H., Akowuah, E.K., Boateng, K.O.: A review of opensource network access control (nac) tools for enterprise educational networks. International Journal of Computer Applications **106**(6) (2014)

[11] Rikhtechi, L., Rafe, V., Rezakhani, A.: Secured access control in security information and event management systems. Journal of Information Systems and Telecommunication **9**(33), 67–78 (2021)

[12] Roy, S., Sharmin, N., Acosta, J.C., Kiekintveld, C., Laszka, A.: Survey and taxonomy of adversarial reconnaissance techniques. ACM Computing Surveys **55**(6), 1–38 (2022)

[13] Syed, N.F., Baig, Z., Ibrahim, A., Valli, C.: Denial of service attack detection through machine learning for the iot. Journal of Information and Telecommunication **4**(4), 482–503 (2020)

Authors

Kalim Qureshi is an Associate Professor of Information Science Department, Kuwait University, Kuwait. His research interests include network parallel distributed computing, thread programming, concurrent algorithms designing, task scheduling, performance measurement and medical imaging. Dr. Qureshi receive his Ph.D and MS degrees from Muroran Institute of Technology, Hokkaido, Japan in (2000, 1997). He published more than 60 journal papers in reputed journals. His email address: kalimuddin.qureshi@ku.edu.kw and kalimuddinqureshi@gmail.com

Mohsen Al-Shamali is a network engineer at Central Bank of Kuwait. He completed his BS in Management Information and currently he completed his MSIT in 2022. His Email address is mohsen.alshamali@ku.edu.kw

Mostafa Abd-El-Barr received his PhD degree from the Department of Electrical and Computer Engineering, University of Toronto, Canada in 1986. He was with the Department of Information Science, College of Computing Sciences and Engineering (CCSE), Kuwait University 2003-2020. He was also an Adjunct Professor with the ECE Department, University of Victoria (UVic), BC, Canada 2009-2020. He is now the Chairman of the Electrical Engineering Department, Badr University in Egypt. His research interests include Information Security, Design and Analysis of Reliable Fault-Tolerant Computer Systems, Computer-Networks-Optimization, Parallel Processing-/Algorithms, Multiple-Valued Logic (MVL) Design Analysis, VLSI System Design, and Digital Systems Testing. He is the author and/or co-author of more than 185 scientific papers published in journals and conference proceedings/symposia. He has three books published (two are translated to the Chinese Language). Professor Abd-El-Barr is a Senior IEEE Member and a member of the International Association of Engineers (IAENG). He is a Senior International Associate Editor of the International Journal of Information Technology Web Engineering and a member of the Editorial of the International Journal of Computer Science and Security (IJCSS). He is also an official IEEE/EAC/ABET evaluator. Dr. Abd-El-Barr is a Certified Professional Engineer in Ontario, Canada.

Maximizing Cyber Resilience through Efficient Vulnerability Prioritization: The WBTS Model

*

Sindhuja Penchala
School of Computing Sciences & Computer Eng.
University of Southern Mississippi
Hattiesburg, US
sindhuja.penchala@usm.edu

Nick Rahimi
School of Computing Sciences & Computer Eng.
University of Southern Mississippi
Hattiesburg, US
nick.rahimi@usm.edu

***Abstract*—In today's digital landscape, cybersecurity demands effective vulnerability management. Our study demonstrates a risk prioritization approach using weighted base scores and vulnerability titles. This method helps organizations evaluate and categorize vulnerabilities based on impact and exploitability, allowing efficient resource allocation to address critical security threats.**

***Index Terms*—Cybersecurity, Vulnerability Management, Cyber threats, Security Threats, Vulnerability Prioritization**

I. Introduction

A vulnerability is essentially a weakness within a system. It is comparable to an unlocked door, allowing a thief to enter a house effortlessly and take whatever they need. Similarly, if there is any vulnerability in a system, then the hacker can easily hack the system, access, and modify any data within it. The number of vulnerabilities has been growing in large number and there is a need to identify them and prioritize based on factors like severity, base score, impact and exploitability[1].This process of identifying and prioritizing is called Vulnerability Prioritization. It is one of the important processes in cybersecurity that involves systematic evaluation. With the ever-increasing count of vulnerabilities discovered daily, it becomes imperative for organizations to prioritize their remediation efforts effectively to mitigate the most critical risks and allocate resources judiciously.

A method for improving vulnerability prioritizing is crucial because the typical vulnerability and patch management backlog currently has over 200,000 issues.[1] Different vulnerabilities have different risk levels, and it is important to treat them with in SLA (Service Level Agreement) breach. To enable fast remediation within SLA limits, vulnerability prioritization comprises the rigorous discovery and rating them based on parameters like Exploitability, Impact, and System Criticality. This procedure guarantees the maintenance of data availability, confidentiality, and integrity within organizations. Prioritizing flaws also helps with the effective use of Patch Management resources, guarantees legal compliance to avoid fines, protects stakeholder's reputation and trust, reduces operational risks like system outages and unauthorized access, and fosters business continuity[2].

Every day, hundreds of new vulnerabilities are adding to the list of NVD, a US government repository. Data about standards-based vulnerability management, represented by the Security Content Automation Protocol (SCAP), is stored by the U.S. government in the NVD[3]. Even with recent advancements, vulnerability prioritization remains an intricate and multidimensional undertaking that demands a nuanced grasp of both technical and organizational factors. Persistent challenges, such as the rise of zero-day exploits, the interconnected nature of contemporary IT ecosystems, and the ever-evolving tactics employed by cyber adversaries, continually put the effectiveness of existing prioritization methodologies to the test.

Fig 1 represents bar plot of the percentage of Common Vulnerabilities and Exposures (CVEs) released between 2000 and 2023, reveals a significant trend. Initially, in the early 2000s, CVE publishing rates were relatively modest, ranging between 1-2 percent . However, percentages have gradually increased over time, with occasional variations. Notably, the curve steepens significantly beginning around 2010, with publishing rates reaching approximately 6-8 percent in the mid-2010s. Subsequently, beginning in 2017, there has been a significant and prolonged increasing trend, with percentages exceeding 10 percent in recent years and peaking in 2023. This graph highlights a large increase in the disclosure of CVEs over time, especially in the recent decade, indicating an intensified attention on detecting and resolving vulnerabilities within software[4].

To overcome the above challenges, this research

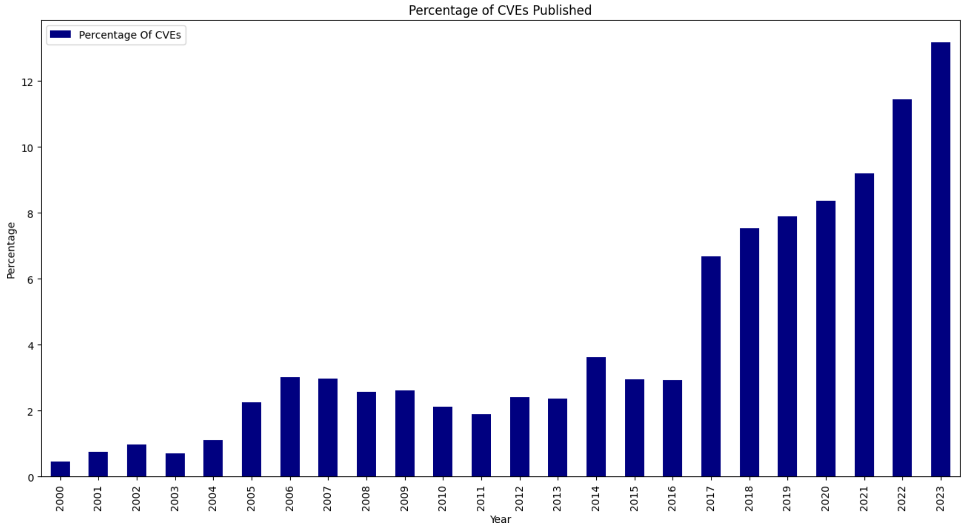


Fig. 1: Percentage of disclosed CVEs per year [4]

paper endeavors to explore the current practices and methodologies employed in vulnerability prioritization.

- It aims to allocate resources to resolve the issues in an organization.
- This research evaluates existing methodologies and proposes new frameworks to improve cybersecurity resilience in a digital society.

There are various sections in this study paper: Section 2 outlines the process of conducting a literature review in relation to vulnerability prioritization; Section 3 provides the methodology; Section 4 gives the data analysis and Section 5 compares the outcomes and supporting metric graphs. Finally, the conclusion and the opportunities for further improvements are covered in Section 6. Here, Table 1 describes the full forms of cyber-related abbreviations.

TABLE I: Abbreviations and their full forms

Abbreviation	Full Form
BSM	Base Score Metrics
CVE	Common Vulnerability Exposures
CVSS	Common Vulnerability Scoring System
CWE	Common Weakness Enumeration
EPSS	Exploit Prediction Scoring System
ESM	Environmental Score Metrics
NVD	National Vulnerability DataBase
OWASP	Open Web Application Security Project
SCAP	Security Content Automation Protocol
SLA	Service Level Agreement
TSM	Temporal Score Metrics
VPR	Vulnerability Priority Rating

II. Literature Review

Vulnerability prioritization is an important part of cybersecurity management because organizations must identify and address the most serious vulnerabilities in their systems to effectively minimize security threats[2,3]. Several research have been undertaken to investigate various approaches and methodologies for vulnerability prioritizing. On average, patch management backlog currently has over 100,000 issues and there are 79.18 CVEs published daily[4].

There are many metrics to resolve this problem. One such method is the Common Vulnerability Scoring System (CVSS) stands out as a popular methodology for this purpose. CVSS can be calculated by using three major metrics: Base Score Metrics (BSM), Temporal Score Metrics (TSM), and Environmental Score Metrics (ESM). While CVSS provides a consistent methodology for assessing vulnerabilities, it has some significant shortcomings. It failed to predict the future threat behavior, instead functioning as a mechanism for researchers and software owners to communicate about responsible disclosure[5].

Several other techniques have been presented to improve CVSS. Exploit Prediction Scoring System (EPSS) predicts the possibility of exploitation within a given timeframe, whereas Vulnerability Priority Rating (VPR) incorporates threat intelligence to represent the current threat landscape[6]. Despite their potential benefits, these approaches may suffer adoption barriers due to issues such as complexity and resource constraints.

In order to overcome these issues, we have proposed a novel approach that can efficiently rank the weaknesses. So that we can allocate the resources to the higher severity issues and remediate with in SLA breach which provides better management and protect the reputation of an organization.

III. Methodology

This section describes the methodology used in our research on vulnerability prioritization, including the methodical approach adopted to ensure a robust and accurate analysis. The procedure includes four

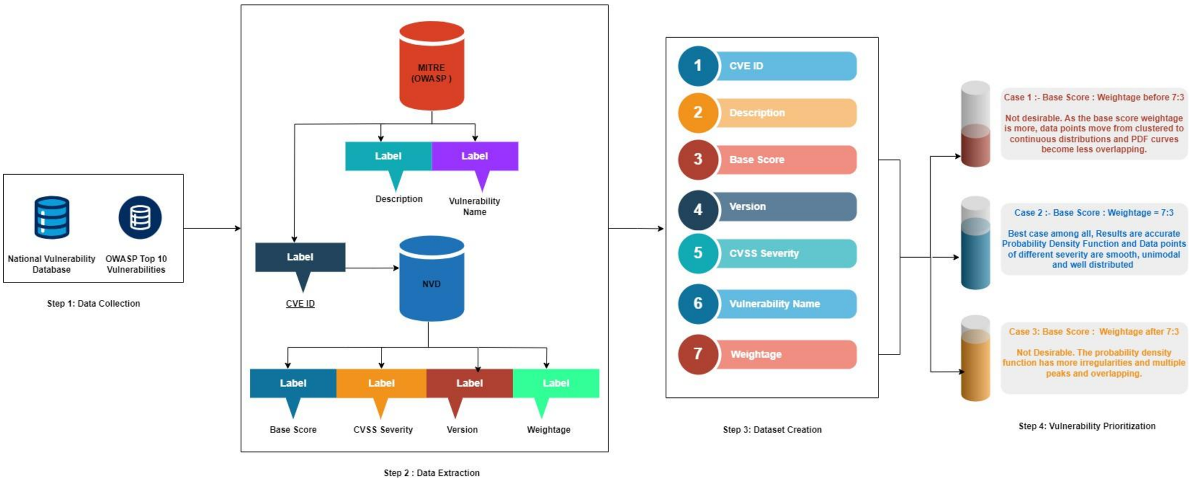

National Vulnerability Database
OWASP Top 10 Vulnerabilities
Step 1: Data Collection
MITRE (OWASP)
Label
Label
Description
Vulnerability Name
Label
CVE ID
NVD
Label
Label
Label
Label
Base Score
CVSS Severity
Version
Weightage
Step 2 : Data Extraction
1 CVE ID
2 Description
3 Base Score
4 Version
5 CVSS Severity
6 Vulnerability Name
7 Weightage
Step 3: Dataset Creation
Case 1 :- Base Score : Weightage before 7:3
Not desirable. As the base score weightage is more, data points move from clustered to continuous distributions and PDF curves become less overlapping.
Case 2 :- Base Score : Weightage = 7:3
Best case among all, Results are accurate Probability Density Function and Data points of different severity are smooth, unimodal and well distributed
Case 3: Base Score : Weightage after 7:3
Not Desirable. The probability density function has more irregularities and multiple peaks and overlapping.
Step 4: Vulnerability Prioritization

Fig. 2: Steps involved in the Methodology

major steps: data gathering, data extraction, creating a dataset, and prioritizing the vulnerability.

3.1 Data Collection

To ensure robust system security, it is crucial to effectively create a dataset. We have collected data from trusted sources, which include the National Vulnerability Database, and the MITRE-Common Weakness Enumeration.

The National Vulnerability Database (NVD) is the US government's repository of standards-based vulnerability management data, managed by the National Institute of Standards and Technology (NIST)[3]. It uses the Common Vulnerabilities and Exposures (CVE) system to catalogue known vulnerabilities, including information such as descriptions of the vulnerability, severity rankings, impact and exploitability. The NVD's centralized vulnerability data enables security teams to immediately identify relevant threats and prioritize repair activities, thereby automating vulnerability management operations.

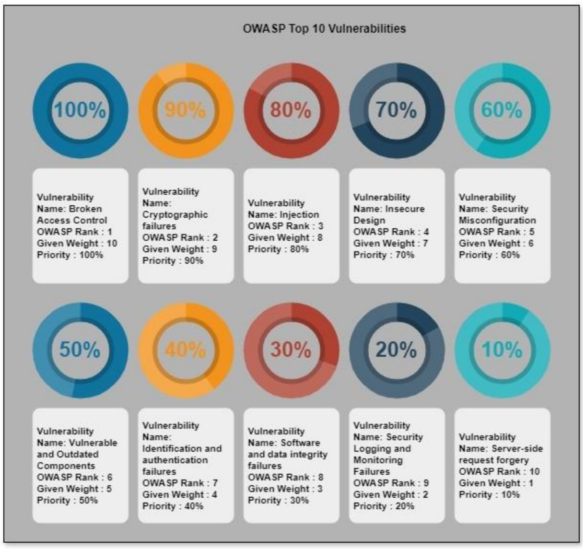
Whereas, cwe.mitre.org is an official website which is managed by MITRE, a US non profit organization. It gives the information about the Common Weakness Enumeration (CWE), a community-developed list of common software and hardware flaws that can lead to vulnerabilities (OWASP top ten). Open Web Application Security Project (OWASP), a non-profit organization dedicated to improving web application security[7]. It has become a crucial resource for developers in the era of cloud-native applications. This list provides up-to-date information on the most significant and widespread vulnerabilities, ranked according to their impact and prevalence. These sources provide valuable information about known vulnerabilities, common attack vectors, and best practices for securing systems and applications.

By utilizing data from NVD and MITRE, organizations can enhance their understanding of potential security threats and take proactive measures to mitigate risks[3,8]. This dataset can be used for security analysis, threat detection, and vulnerability management to safeguard systems and networks from potential cyber-attacks. By analyzing the data, businesses can proactively address any weaknesses and implement necessary security patches and updates to fortify their defenses. Automation of security measurement, compliance, and vulnerability management is made possible by this data. Databases containing software vulnerabilities connected to security, product names, impact metrics, and security checklist references are all included in the NVD[9].

3.2. Data Extraction

We obtained information from cwe.mitre.org and NVD using a technique known as web scraping. Web scraping technologies can help organizations automate the process of acquiring up-to-date information on known vulnerabilities, common attack paths, and best practices for system security. It also makes it easier to grow and manage the dataset over time, ensuring that businesses always have access to current, relevant information about emerging threats. By incorporating web scraping, firms can better uncover vulnerabilities, stay ahead of potential security threats, and take proactive steps to improve system security. It can be done using a Python package called Beautiful Soup. Beautiful Soup, with its ability to parse HTML and XML documents, makes web scraping easier by offering tools to browse document structure and extract required data.

In fig 2, step 2 represents the process of creating dataset from CWE – OWASP and NVD databases. The MITRE-CWE cite contains data about the software weaknesses which includes CVE ID, Description and name of the vulnerability. Whereas, in NVD we can find the data of CVE ID, Summary, CVSS Severity and Base Score. Both URLs contain CVE ID in common. So once the data from CWE-OWASP is extracted, we used the attribute CVE ID to retrieve the required columns from NVD. In this way, we have created the dataset in the form a csv file. Here, we have added one



Vulnerability Name: Broken Access Control
OWASP Rank : 1
Given Weight : 10
Priority : 100%

Vulnerability Name: Cryptographic failures
OWASP Rank : 2
Given Weight : 9
Priority : 90%

Vulnerability Name: Injection
OWASP Rank : 3
Given Weight : 8
Priority : 80%

Vulnerability Name: Insecure Design
OWASP Rank : 4
Given Weight : 7
Priority : 70%

Vulnerability Name: Security Misconfiguration
OWASP Rank : 5
Given Weight : 6
Priority : 60%

Vulnerability Name: Vulnerable and Outdated Components
OWASP Rank : 6
Given Weight : 5
Priority : 50%

Vulnerability Name: Identification and authentication failures
OWASP Rank : 7
Given Weight : 4
Priority : 40%

Vulnerability Name: Software and data integrity failures
OWASP Rank : 8
Given Weight : 3
Priority : 30%

Vulnerability Name: Security Logging and Monitoring Failures
OWASP Rank : 9
Given Weight : 2
Priority : 20%

Vulnerability Name: Server-side request forgery
OWASP Rank : 10
Given Weight : 1
Priority : 10%

Fig. 3: OWASP Top 10 Vulnerabilities

more column where we assigned weightage to type of vulnerability. Over all, dataset consists of 2211 records with seven columns.

3.3. Dataset Creation

The dataset is created in the form of csv file. It consists of 7 columns which include CVE-id, Description, Base Score, Severity, Version, Vulnerability Name and Assigned weightage.

The fig.3 presents the OWASP Top 10 Vulnerabilities, a widely recognized list of critical security risks in web applications. The rankings are given by the experts based on their impact and likelihood. The lower the rank, the higher the weight is assigned. In the above figure, we can find that Broken access control vulnerability is ranked number 1. It means high weightage should be given with a weight of 10. Then weight 9 is given to the vulnerability Cryptographic Failures as it is ranked second and so on.In this way we have assigned weightage to title and created the seventh column.

3.4. Prioritizing Vulnerabilities

In this stage, we have prioritized vulnerabilities based on the ratio of base score and assigned weights. We have provided with three cases to illustrate the desired outcome:

Case 1 (Base Score : Weight before 7:3): As the base score's influence increases, data points transition from clustered to continuous distributions, and PDF curves become less overlapping and more distinct.

Case 2 (Base Score : Weightage = 7:3): This is considered the best case, where the probability density function and data points of different severity are smooth, unimodal, and well-distributed.

Case 3 (Base Score : Weightage after 7:3): Not desirable. Not desirable. The probability density curve has more abnormalities, multiple peaks, and overlappings.

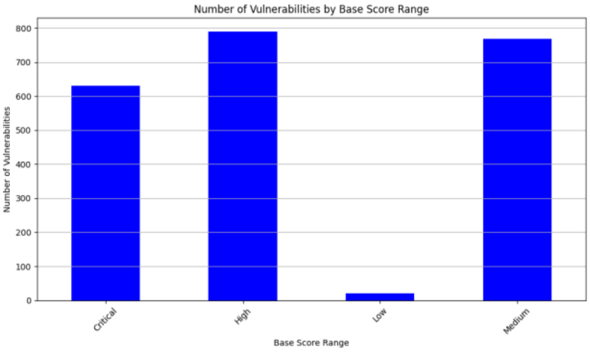


Fig. 4: Bar chart representing Number of different vulnerabilities

The methodology aims to find an optimal combination of base score and weight that results in a well-distributed and unimodal probability density function, allowing for effective prioritization of vulnerabilities based on their severity and importance.

IV. Data Analysis

4.1. Data Visualization

We have visualized data in two forms:bar plot and a pie chart. These graphs helps us to easily identify the number of vulnerabilities taken in each category like low, medium, high and critical.

4.1.1. Bar Plot Visualization

The figure 4 depicts a bar chart with the amount of vulnerabilities classified by various base score ranges. Base score ranges include ”Critical,” ”High,” ”Low,” and ”Medium.” The y-axis reflects the number of vulnerabilities, and the x-axis depicts the various base score ranges. The graphic shows that the ”High” base score range has the most vulnerabilities, followed by ”Critical,” ”Medium,” and ”Low” categories, respectively. This graphic depiction enables a quick comparison of the vulnerability distribution across various severity levels, as indicated by the base score ranges.

4.1.2. Pie chart Visualization

Figure 5 shows a pie chart that depicts the distribution of vulnerabilities across various base score ranges. The greatest chunk, indicated in green, falls inside the ”High” base score range, accounting for 35.7 percent of the vulnerabilities. The second-largest section, depicted in light blue, reflects the ”Medium” base score range, which includes 34.8 percent of the vulnerabilities. The orange slice represents the ”Critical” base score range, which accounts for 28.6 percent of all vulnerabilities. The smallest section, highlighted in red, reflects the ”Low” base score range, which includes only 1 percent of the vulnerabilities. This graphic depiction provides a clear overview of the relative proportions of vulnerabilities classified by severity levels using the base score ranges.

4.2. Metrics involved for prioritizing vulnerabilities

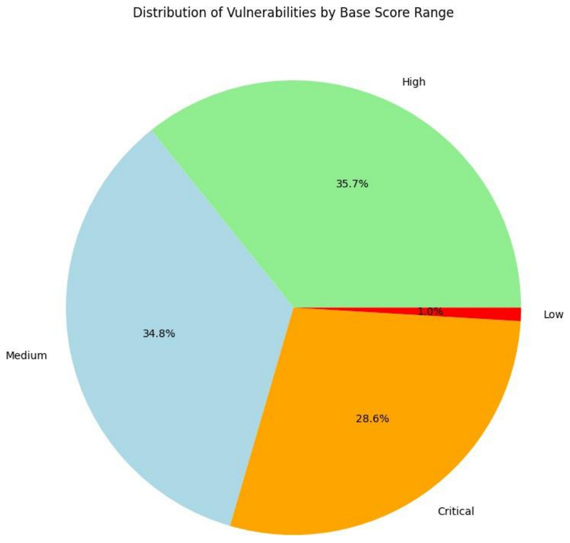


Fig. 5: Pie chart representing Number of different vulnerabilities

There have been many methods used to rank the CVEs. One of the common and popular methods is CVSS. The Common Vulnerability Scoring System is one such technique that is frequently used for vulnerability rating (CVSS)[10]. A standardized framework called CVSS is used to evaluate the severity of security flaws. It offers a quantitative method to rank flaws according to their attributes. CVSS is calculated based on three components: Base, Temporal and Environmental metrics. We have proposed a new metric which prioritizes vulnerabilities based on weights assigned to base score and title of a CVE. The names are taken from cwe.mitre.org top 10 OWASP vulnerabilities. The metrics we used are Base score and weighted title of vulnerability.

4.2.1 Base Metrics

The CVSS Base Score measures the vulnerability's intrinsic severity, regardless of time or environment. It is estimated using the vulnerability's properties, such as attack vector, attack complexity, privileges required, user interaction, scope, confidentiality impact, integrity impact, and availability impact[11]. These measurements remain constant over time and are unaffected by real-world exploitability or compensating mechanisms established by an organization. The Base Score ranges between 0 and 10, with higher ratings suggesting more serious vulnerabilities.

The function can be represented as:

$$f(\text{ISS}) = \begin{cases} 0 & \text{if ISS} \leq 0 \\ \text{SU} & \text{if ISS} > 0 \text{ and SU} \\ \text{SC} & \text{if ISS} > 0 \text{ and SC} \end{cases} \quad (1)$$

$$\text{SU} = \text{Round}(\text{Minimum}[(\text{I} + \text{E}), 10])$$

$$\text{SC} = \text{Round}(\text{Minimum}[1.08 \times (\text{I} + \text{E}), 10])$$

Here, ISS represents Impact Sub Score, SU represents Scope Unchanged, SC means Scope Changed, I represents Impact and e means Exploitability.

a. Exploitability: Exploitability Metrics are important in vulnerability assessment because they focus on the underlying properties of the vulnerability rather than specific configurations or compensating mechanisms. These metrics, which consist of four components - Attack Vector, Attack Complexity, Privileges Required, and User Interaction - help to assess the exploitability of the vulnerability[12].

• The Attack Vector shows an attacker's level of access, which can be Network, Adjacent, Local, or Physical.

• Attack Complexity distinguishes between the ease of exploitation, which is evaluated as Low or High.

• Privileges Required defines the level of access required for effective exploitation, which is classified as None, Low, or High.

• User Interaction classifies whether user interaction, other than that of the attacker, is required for the exploit to succeed, with None and Required options.

The exploitability is calculated using below formula,

The formula for Exploit can be represented as:

$$\text{Exploit} = 8.22 \times \text{AttrackVector} \times \text{AttrackComplextiy} \times \text{PrivilegeRequired} \times \text{UserInteraction} \quad (2)$$

b. Impact: Impact Metrics used in vulnerability assessments analyze the effects on the impacted system's CIA Triad (Confidentiality, Integrity, and Availability)[13].

• Confidentiality evaluates the extent of sensitive information leakage, classifying it as High, Low, or None based on the attacker's access level.

• Integrity assesses the potential alteration of protected data, with values ranging from None to High depending on the level of tampering permitted by the vulnerability.

• Availability governs the accessibility of information after exploitation, with values ranging from None to High reflecting the level of unavailability or service interruption. .

The formula for impact for different scope

The formula for impact for different scopes can be represented as:

$$\text{Impact for Scope Unchanged} = 6.42 \times \text{ISCBase} \quad (3)$$

$$\text{Impact for Scope Changed} = 7.52 \times [\text{ISCBase} - 0.029] - 3.25 \times [\text{ISCBase} - 0.02]^{15} \quad (4)$$

Where:

$$\text{ISCBase} = 1 - [(1 - \text{ImpactConf}) \times (1 - \text{ImpactInteg}) \times (1 - \text{ImpactAvail})]$$

c. Scope: The scope is another parameter in vulnerability assessment that is used to calculate the base score. It determines the potential impact of a vulnerability beyond its immediate surroundings. It determines if a vulnerability in one system or component can propagate to other interconnected systems or components. The Scope metric aids companies in understanding the scope of a vulnerability's impact, allowing them to assess the possible magnitude of damage and execute suitable mitigation measures. It is divided into two ratings: Changed and Unchanged[14].

• A Changed rating implies that the exploited vulnerability may have cascading consequences on other systems or components beyond its security scope.

• An unchanged rating shows that the damage is limited to the local security authority.

4.2.2 Vulnerability Title:

The second metric used in this research is the name of the vulnerability. We collected OWASP top 10 vulnerabilities which have been ranked by the experts based on data factors provided by few organizations. The current Top 10 list is more driven by data analysis than previous versions, but not excessively so. Eight of the ten categories were chosen directly from the supplied data, while the other two came from high-level results of the Top 10 community survey. They have listed few weaknesses like broken access control , cryptographic failures, Injection which became the most serious threats in present world[15].

The formula of our proposed methodology is:

$$\text{Priority Score} = W1 \times \text{Base Score} + W2 \times \text{Weighted Score} \quad (5)$$

Where:

$W1$ = Weight given to Base Score (70%)

$W2$ = Weight given to Weighted Score (30%)

The table 2 shows the names of most important risks with ranking and weight assigned to them. The rankings are provided by cwe.mitre.org for OWASP top 10 vulnerabilities by the experts based on some data factors and a survey. The lower the rank the higher the weight is given because high priority issues should be resolved first. So, when more weights are provided then priority score gets increased and thus, we prioritize them for remediation.

V. COMPARISON AND RESULTS

In this section, we have achieved a probability density function and data point representation graphs using the priority score.

5.2. Case 1: Base Score : Weight before 7:3

TABLE II: Weights Assigned to CVEs

Name of the Vulnerability	Rank	Weight- Score
Broken Access Control	1	10
Cryptographic Failures	2	9
Injection	3	8
Cross Site Scripting (XSS)	4	7
Insecure Design	5	6
Security Misconfiguration	6	5
Identification and Authentication Failures	7	4
Software and Data Integrity Failures	8	3
Security Logging and Monitoring Failures	9	2
Server Side Request Forgery (SSRF)	10	1

Table 3 represents the graphs for the Datapoint representation and probability distribution function, drawn by taking various ratios of the base score and assigned weights. From the graphs, 6(a) to 6(g), we analysed the progression across all seven graphs (from 0:1 to 6:4 ratio of base score to assigned title score), we see a definite trend in vulnerability prioritization:

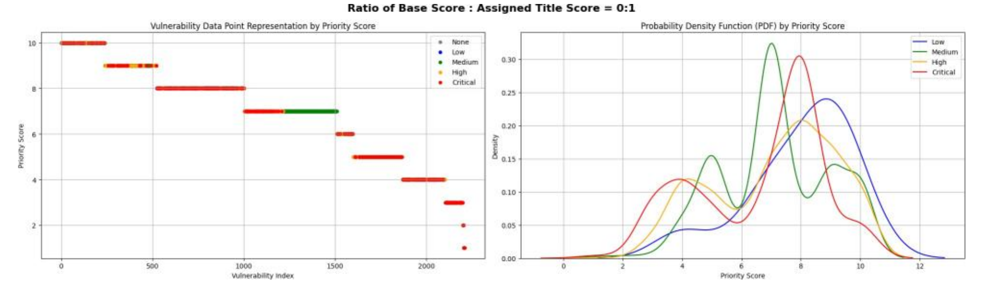
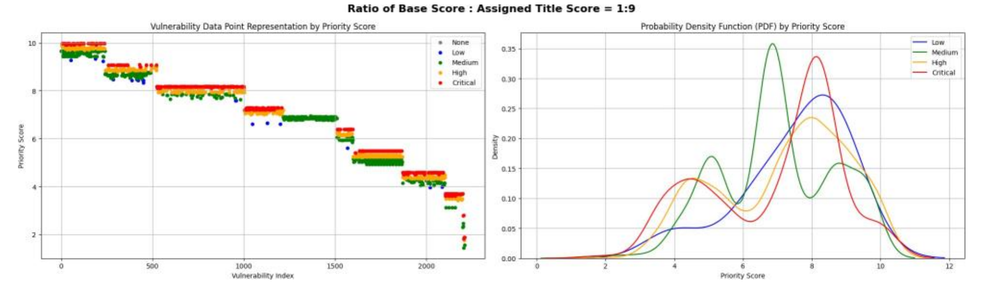
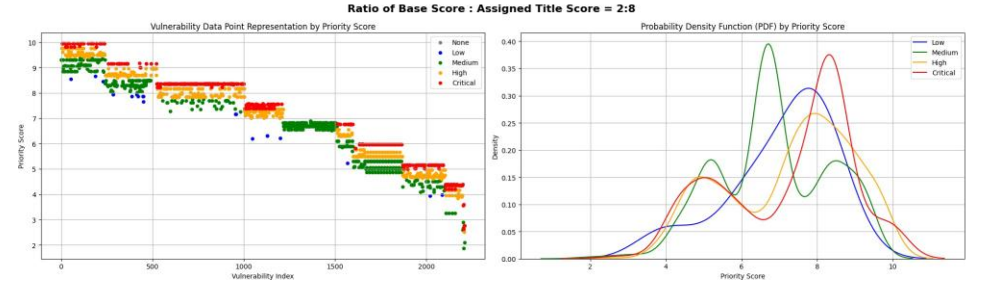
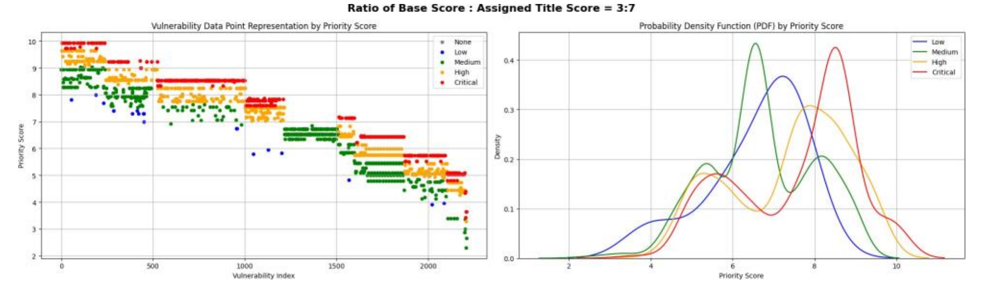
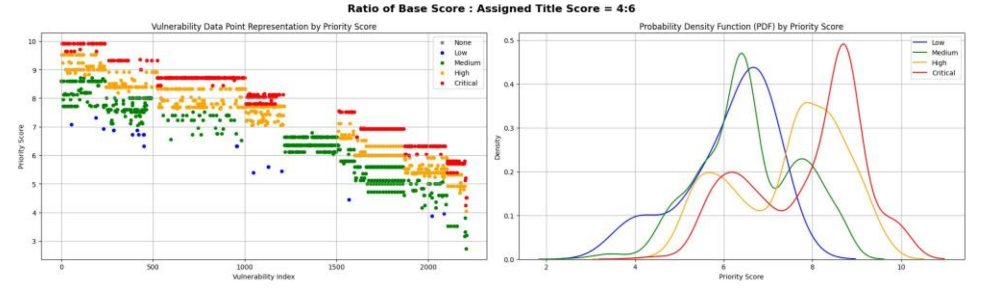
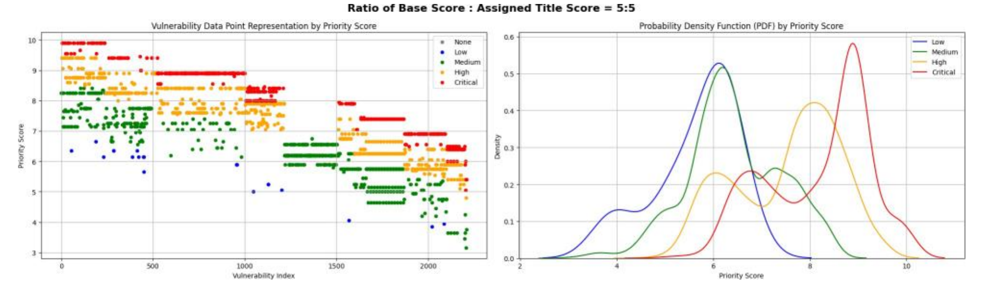
As the base score's effect grows, there is a visible shift from extremely discrete, clustered data points to a more continuous, spread-out distribution. For each priority level, the PDF curves shift from multi-peaked and overlapping to more distinct, separated curves. In the 0:1 and 1:9 ratios, vulnerabilities are classified rigidly, but the 3:7 and 4:6 ratios strike a balance between categorization and nuanced grading. The 5:5 and 6:4 ratios show a sharper distinction of priority levels, particularly for major vulnerabilities, as the PDF curves become more apparent and less overlapping.

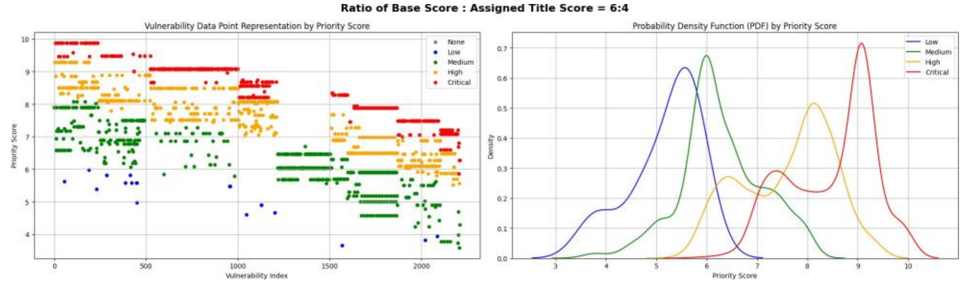
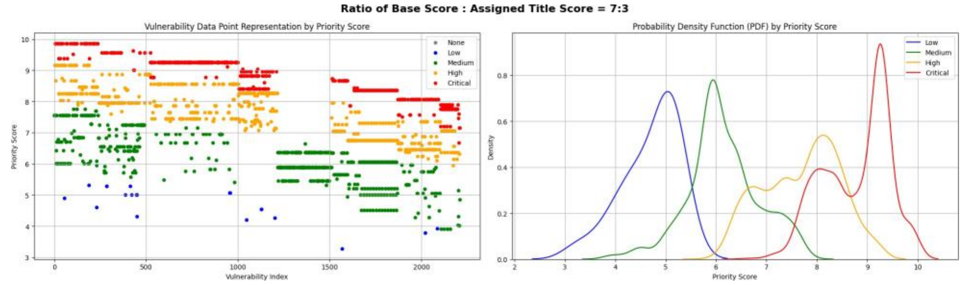
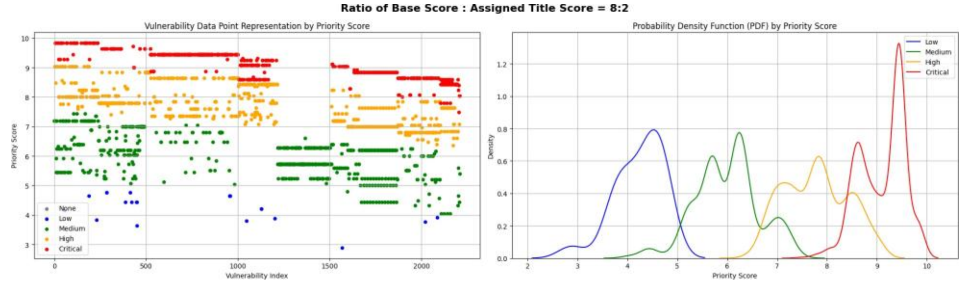
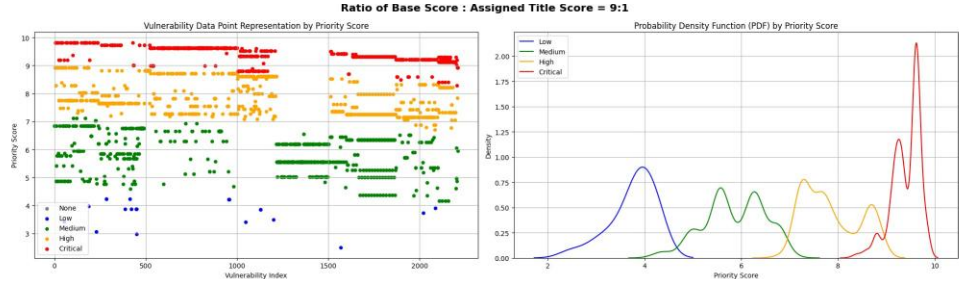
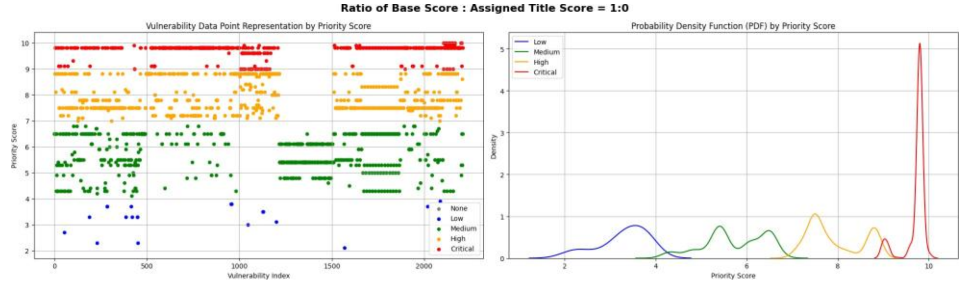
5.1. Case 2: Base Score Range: Weight = 7:3

This is the best ratio among all the cases. Here, highlighted blue colour image depicts a graph of the Data point representation and Probability Density Function (PDF) of priority scores with the ratio of 7:3. The left side of fig. 6(h) shows the representation of data point of the weaknesses from index 0 which is on X axis, whereas Y axis represents the priority score which ranges from 3 to 10. The red data points represent critical vulnerabilities, orange represents high, green shows medium and blue means low priority weakness. There are some orange dots above the range of 9 which has to be given priority compared to the red points which are below the score 9. Even though orange indicate high, they have to be resolved first as they have more priority score compared to few critical vulnerabilities.

The right side of graph 6(h) features four curves, each reflecting a different base score range: low, medium, high, and critical. The x-axis indicates the priority score, which ranges from 2 to 10. The y-axis shows the density, or probability occurring at a specific priority score. The Low base score range has a bell-shaped curve that peaks at a priority score of 2, indicating that low base score ranges are more likely to have a priority score of 2, where the medium base score range peaks at a priority score of 4, the High base score range has a curve with a peak near a priority

TABLE III: Graphs showing the relationship between the base score and assigned weights

Assigned Weight	Graph
Case 1	
6(a) - 0:1	 Ratio of Base Score : Assigned Title Score = 0:1
6(b) - 1:9	 Ratio of Base Score : Assigned Title Score = 1:9
6(c) - 2:8	 Ratio of Base Score : Assigned Title Score = 2:8
6(d) - 3:7	 Ratio of Base Score : Assigned Title Score = 3:7
6(e) - 4:6	 Ratio of Base Score : Assigned Title Score = 4:6
6(f) - 5:5	 Ratio of Base Score : Assigned Title Score = 5:5

Assigned Weight	Graph
6(g) - 6:4	
Case 2	
6(h) - 7:3	
Case 3	
6(i) - 8:2	
6(j) - 9:1	
6(k) - 1:0	

score of 6 and critical score peaks at a score of 9.

5.3. Case 3: Base Score Range: Weight after 7:3

Figures 6 (i), 6 (j), and 6 (k) compare different base scores to assigned title score ratios (8: 2, 9: 1 and 1: 0, revealing a consistent development in vulnerability classification and distribution. As the ratio favors the base score (from 8:2 to 1:0), there is a noticeable movement toward higher priority scores, particularly for serious vulnerabilities, with Probability Density Function (PDF) curves getting taller and narrower, indicating a more concentrated distribution. The 7:3 ratio stands out as the best balanced method, providing visible separation between priority levels without over-polarization as shown in the 9:1 and 1:0 ratios. This balance enables nuanced prioritization by capturing both the technical severity of the base score and the contextual importance of the assigned title score, resulting in a practical and successful technique for addressing vulnerabilities in complex IT infrastructures.

Overall, the analysis demonstrates that as the base score's weight grows, vulnerability prioritizing switches from a balanced, nuanced approach to a more rigid, technically driven classification. Ratios like 5:5, 6:4, and 7:3 find the optimum balance, providing obvious boundaries across priority levels while allowing flexibility within each category. These ratios efficiently integrate technical severity and contextual relevance, resulting in more precise and useful vulnerability assessments. The 7:3 ratio is shown to be the most successful, providing a well-balanced strategy that provides both technical correctness and contextual relevance in ranking.

VI. Conclusion

In conclusion, organizations must prioritize the resolution of the most serious threats by carefully reviewing and ranking security vulnerabilities. We have proposed an approach based on their weighted base title score. In this study, we have achieved the best results for the ratio of Base score to Title 7:3. It helps in ranking the critical CVE's, by not only using the CVSS base score but also based on the name of the top ten vulnerabilities. This technique helps to reduce the risk of security breaches and their possible impact on corporate operations. Implementing a structured vulnerability prioritization approach allows companies to make more informed decisions about resource allocation and risk management. It enables them to proactively address the most important security risks, improving their overall security posture.

In the future, we will compare the different ways in which machine learning models are used to effectively automate the procedure. We also create a chatbot that provides specific information such as prevention measures, assaults, and the reasons for the top ten vulnerabilities. The chatbot can be used as a quick and easy instructional tool to help people learn common vulnerabilities, their implications, and mitigation measures. By providing fast access to current information, the chatbot can help with proactive security measures, incident response, and vulnerability prioritization. This tool is especially valuable for developers, security teams, and businesses, allowing them to stay informed and take proper precautions to secure their systems.

References

[1] Farris, Katheryn A., et al. ”Vulcon: A system for vulnerability prioritization, mitigation, and management.” ACM Transactions on Privacy and Security (TOPS) 21.4 (2018): 1-28.

[2] J. Yadav, Geeta, et al. ”SmartPatch: A patch prioritization framework.” Computers in Industry 137 (2022): 103595.

[3] https://nvd.nist.gov/vuln

[4] https://www.jerrygamblin.com/

[5] Bulut, Muhammed Fatih, et al. ”Vulnerability prioritization: An offensive security approach.” arXiv preprint arXiv:2206.11182 (2022).

[6] Hughes, Chris, and Nikki Robinson. ”Vulnerability Scoring and Software Identification.” (2024): 79-114.

[7] https://www.clouddefense.ai/owasp-top-10-vulnerabilities/

[8] CWE - CWE-1344: Weaknesses in OWASP Top Ten (2021) (4.14) (mitre.org)

[9] Hore, Soumyadeep, Ankit Shah, and Nathaniel D. Bastian. ”Deep VULMAN: A deep reinforcement learning-enabled cyber vulnerability management framework.” Expert Systems with Applications 221 (2023): 119734.

[10] Jung, Bill, Yan Li, and Tamir Bechor. ”CAVP: A context-aware vulnerability prioritization model.” Computers & Security 116 (2022): 102639.

[11] Sharma, Abhishek, Sangeeta Sabharwal, and Sushama Nagpal. ”A hybrid scoring system for prioritization of software vulnerabilities.” Computers & Security 129 (2023): 103256.

[12] Elder, Sarah, et al. ”A Survey on Software Vulnerability Exploitability Assessment.” ACM Computing Surveys 56.8 (2024): 1-41.

[13] Dodiya, Bindu, Umesh Kumar Singh, and Vivaan Gupta. ”Trend analysis of the CVE classes across CVSS metrics.” International Journal of Computer Applications 975 (2021): 8887.

[14] Costa, Joana Cabral, et al. ”Predicting CVSS metric via description interpretation.” IEEE Access 10 (2022): 59125-59134.

[15] Aljabri, Malak, et al. ”Testing and exploiting tools to improve owasp top ten security vulnerabilities detection.” 2022 14th International Conference on Computational Intelligence and Communication Networks (CICN). IEEE, 2022.

Image Processing Without Sacrificing the Functional Paradigm

Antoine Bossard [*]
Kanagawa University, Yokohama, Kanagawa 221-8686, Japan

Abstract

Functional programming can be called modern programming in that it enables robust development and favours the programmer over hardware and performance considerations. Some characteristic features of the functional paradigm, such as map, have indeed been introduced into languages of other programming paradigms, like C++ and JavaScript. In spite of this deserved success, some challenges remain, such as input-output (I/O) operations, which often involve compromises with respect to the functional model. For instance, heavy memory I/O applications may keep prospective users afar. In this paper, we give a constructive proof of the practicability of the functional paradigm for such a scenario, by concretely considering image processing with the functional programming language Racket (a Lisp dialect). Both theoretical and experimental quantitative evaluations are conducted to show the performance of the implemented algorithms. Furthermore, in an attempt at establishing the capabilities and versatility of functional programming, this work also covers parallel processing, on both a single core and multiple cores.

Key Words: graphics; dithering; parallel; programming; software; Racket.

1 Introduction

The importance of the functional paradigm and functional programming need not be proven any more [6]. Functional constructs have even penetrated other paradigms, such as the object-oriented one with, for example, JavaScript's `Array.prototype.map` method [13] and C++'s `std::apply` function [7]. This is also the case of anonymous functions [14]. Moreover, it has been shown that functional programming is positive for Internet of Things (IoT) applications [5]. Considering concrete implementations of the functional paradigm, Lisp dialects can be found in robotics [11] and microcontroller applications [8]. It is thus no wonder that the functional paradigm is increasingly popular [2].

Despite all the advantages of the functional paradigm, users are likely to rapidly face some challenges, not to say limitations, inherent to this programming model. They mostly concern input-output (I/O) operations as those generally harm referential transparency [10].

To address some of these issues, several solutions have been proposed. For instance, Haskell, another functional programming language, relies on monads [9]. In addition, the SOF programming paradigm has been described to address the state tracking issue of pure functional programs, and especially those featuring lazy evaluation [1]. The Racket programming language is said multi-paradigm in that it allows imperative programming, albeit clearly stating that it is a best practice to limit as much as possible such non-functional constructs [3]. High-performance image processing with a functional programming style approach was proposed by Se´rot *et al.* [12], but it requires a specific, complex hardware architecture.

Our objective in this paper is to give a constructive proof that shows the functional programming paradigm remains practicable even when conducting image processing, that is, memory operations on rather large data. In other words, that such heavy I/O operations are not an excuse to sweep the functional paradigm aside [4].

To this end, we have selected the Racket language, based on Scheme and thus a Lisp dialect, as it tolerates imperative programming when needed, as previously recalled [3], which drastically improves usability over pure functional languages like Haskell.

The rest of this paper is organised as follows. First, preliminaries for this work are presented in Section 2. Then, a first image processing algorithm, dynamic palette calculation, is discussed in Section 3. Next, this algorithm is reused to describe in Section 4 a dithering algorithm. Dithering is further discussed from the parallel processing point of view in Section 5. Finally, concluding remarks are made in Section 6.

2 Preliminaries

About the notations used hereinafter, for the sake of brevity but without introducing any ambiguity, set operations such as $\setminus$ (exclusion) and $|\ldots|$ (cardinality) are sometimes applied to lists (i.e. sequences).

Regarding image file loading and pixel access, Racket provides convenient functions. First, `read-bitmap` takes a path to an image file (the GIF, JPEG, PNG, BMP and a few other formats are supported) and instantiates and returns the corresponding `bitmap%` object. Second, the pixel values, that is colours, that make the bitmap can be retrieved with the `get-argb-pixels` method of the `bitmap%`

[*] Graduate School of Science. Email: abossard@kanagawa-u.ac.jp.

class; note that memory needs to be allocated beforehand for their storage. Conversely, the pixels of a `bitmap%` object can be set with the `set-argb-pixels` method. Finally, a `bitmap%` object can be saved to an image file with the `save-file` method, which is how we produced the images included hereafter.

A bitmap can be conveniently displayed inside a window thanks to Racket's GUI components: it suffices to attach a `canvas%` object to a `frame%` object and to call the `draw-bitmap` method of the `dc%` class inside the paint callback of the canvas.

Finally, we have noticed that relying on the class `color%` provided by Racket to represent a colour is better avoided for two reasons: first, it significantly slows algorithm implementations down compared to, for instance, a simple triple (list), and second, it forbids using non-integer decimal values for RGB channel values. Non-integer decimal values are desirable for error diffusion as considered later.

3 Dynamic Palette Calculation

Two palette calculation methods plus one performance optimization solution are described in this section.

3.1 Main Approach

Be it for compression purposes or to adhere to a standard like the Graphics Interchange Format (GIF), applying a colour palette to an image has always been an essential issue of computer graphics. We thus start by considering this well-known scenario to realise our constructive proof of the relevance and practicability of the functional paradigm for image processing.

Although simple and fast, relying on a static colour palette, such as IBM's 16-colour CGA palette, produces below par results. So, given that modern computer hardware allows it, it is instead wiser to dynamically calculate an optimised palette from the image that is to be rendered.

We proceed as follows: first, we enumerate the different colours used in the image, and for each of them, we record their frequency (i.e. the number of times the colour is used in the image). This can be implemented simply: consider the list l of all the pixels (i.e. colours) making the image; get the first colour of l, say c, count in l the number n_c of pixels of same colour c and repeat this process from the list of colours that differ from c, list which becomes the new l. This can be easily realised with the `partition` function called in a way that it returns the list of colours equal to c, thus inducing n_c, and the list of colours different from c, thus inducing the new list l.

Now that all the image colours and their frequency have been obtained, the next task is to retain as many colours as can hold the palette, say k. A na¨ıve approach to this issue is to sort the obtained list of colours in descending order of frequencies and to copy the first k colours into the palette. (All the colours of the image are retained if there are less than or the same number as the palette size k.) This way, the palette consists of the most frequent colours of the image. Although simple, this first dynamic palette calculation method produces unsatisfactory results. A picture is worth a thousand words: refer to Figure 1a.

So, instead of selecting the colours to be retained inside the palette depending on their respective frequencies, it is indeed better to group colours according to their similarity: for two similar colours, the one with the higher frequency is retained, the other discarded. Precisely, we start by sorting the image colours in ascending order of their frequency so that the most infrequent colours will be grouped, that is eliminated, first; say this is the colour list l. Then, we iterate l, starting with its first colour, say c, each time finding within $l \setminus \{c\}$ the colour that is the nearest to c, say c', and we retain from c and c' only the one with the higher frequency, as explained. This is repeated with the new, smaller sorted list of colours $l \setminus \{\tilde{c}\}$, with $\tilde{c} \in \{c, c'\}$ the discarded colour. This iteration is terminated as soon as the number of the remaining colours, that is $|l|$ the size of l, is smaller than or equal to the palette size. The superiority of this second dynamic palette calculation method is clear: refer to Figure 1b.

Finally, a word on palette application to an image: in one single pass, for each pixel (colour) of the image, iterate the palette to find the nearest colour, which is stored as the new pixel value. The nearest colour is found by simply summing the difference between each of the three RGB channels.

3.2 Optimization

If instead of relying, as previously, on a simple list to count distinct colours we rely on a hash table, performance can be raised since Racket provides a hash table mechanism with constant time access operations. Furthermore, hash tables can be immutable, which allows us to avoid any trade-off with the functional paradigm for that matter.

Concretely, we iterate the image pixels only once (i.e. in a single pass), each time incrementing the hash value corresponding to the pixel colour; colours serve as hash keys.

Source code is given in Listing 1 to illustrate the elegance of this optimised approach which induces significantly higher performances (refer to Section 3.3).

Listing 1: Counting colours faster, in one pass, with a hash table.

```
(hash->list ; returns colours and frequencies conventionally as a list
  (foldl (lambda (c hash-table) ; 'c' is the current colour
           (let ([current-value (hash-ref hash-table c 0)])
             (hash-set hash-table c (add1 current-value))))
    (hash) image-colours)) ; '(hash)' returns a new, empty hash table
```

3.3 Quantitative Evaluation

We begin by considering the worst-case time complexity of the non-optimised approach to colour enumeration and frequency calculation (i.e. based on the `partition` function). In the worst case, which corresponds to an image with no two pixels of the same colour, for each of the n pixels of the image, the remaining pixels are split into pixels of the same colour (none in the worst case as just explained) and pixels of a different colour. This process is repeated for each pixel, each time starting over from

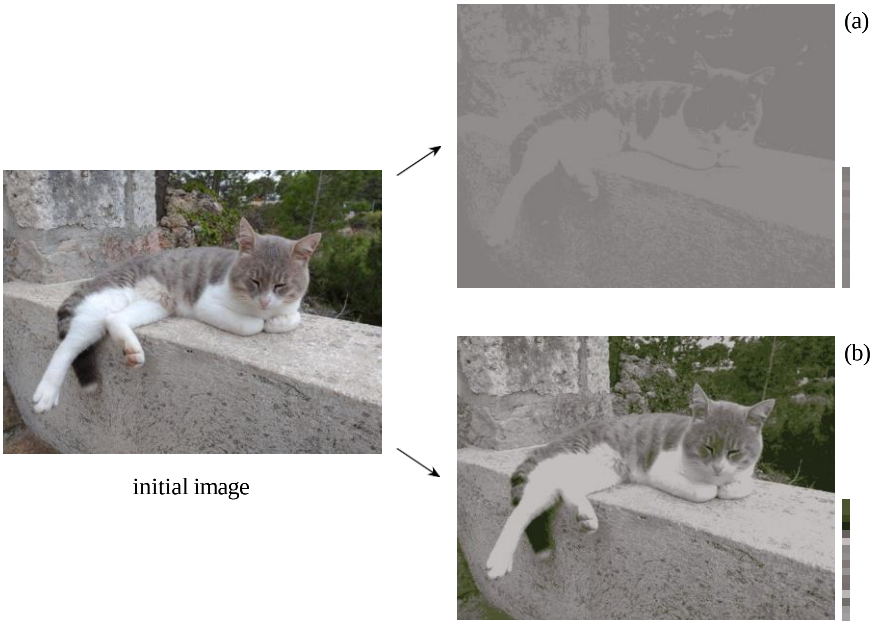


Figure 1: Dynamic palette calculation and application to a sample photograph. The 16-colour palette is shown beside the picture. (a) Na¨ıve dynamic palette calculation. (b) Improved dynamic palette calculation. (Photograph taken by the author.)

the remaining pixels (i.e. the pixels not classified yet). Hence, the worst-case time complexity of this colour enumeration and frequency calculation is $O(nk) = O(n^2)$ with k the number of distinct colours in the image.

On the other hand, the optimised approach for colour enumeration and frequency calculation based on a hash table is faster: since hash table access operations are constant time $O(1)$, one single pass of the n image pixels suffices, thus inducing an $O(n)$ time complexity.

Next, the enumerated colours need to be grouped. So, in either approach, palette calculation requires an additional time complexity of $O(\log(k/p) \times k \log k)$, with k the number of distinct colours in the image and p the maximum number of colours inside the palette. Indeed, since each iteration of the k enumerated image colours eliminates at least one and at most $k/2$ colours, a palette of size p will be obtained after $\log(k/p)$ iterations. So, the k enumerated image colours are sorted no more than $\log(k/p)$ times, which induces the dominant time complexity.

As a result, in the worst case (i.e. $k = n$), the non-optimised approach requires $O(n^2 + n \log n \log(n/p))$ and the optimised one $O(n \log n \log(n/p))$. Moreover, we have empirically confirmed the theoretically established worst-case time complexity with a computer experiment: we have run an implementation of the described dynamic palette calculation algorithm for several image files as follows. We have selected one photograph, so that numerous colours be included and palette calculation be thus meaningful, and we have resized it to produce several other image files of lower resolutions. The different image resolutions have thus enabled us to vary the value of n.

Next, we make a remark regarding the number of colours k. While variations of k in the image files selected for this experiment could impact the measurement of the average time complexity, in the worst case k equals n, a case which has been considered when establishing the worst-case time complexity above. So, variations of k will not prevent experimentally confirming the theoretically established worst-case time complexity. In practice, the original photograph, likely because of physical limitations of the camera sensor or the camera image compression algorithm, may have a lower k/n ratio, and even a lower k, than after applying a first resizing operation. So, in an attempt to stabilise the k/n ratio and thus to estimate the average time complexity by emphasising the variations of n, we consider only images resulting from at least one resizing operation. Hence, the original photograph is not used in this experiment other than to produce the experiment images by resizing.

This experiment has been conducted on a computer running the Debian GNU/Linux 12 (64-bit) OS equipped with a 12th generation Intel Core i5-12400 processor and 16 GB RAM. The experimental results show the difference between the non-optimised dynamic palette calculation method, based on the `partition` function, and the optimised dynamic palette calculation method, based on a hash table. Time measurements were reported by the `time` function of Racket (applied to the dynamic palette calculation function), whose “real time” value was retained. The photograph of Figure 1 before applying a palette has been considered in the following different resolutions (in pixels): 591 × 443 (i.e. $n = 261813$), 443 × 332 (i.e. $n = 147076$), 296 × 222 (i.e. $n = 65712$) and 148 × 111 (i.e. $n = 16428$). The number of colours k was 29 721, 23 190, 16 326 and 6 854, respectively. The palette size p was fixed to 16. The obtained results are illustrated in Figure 2.

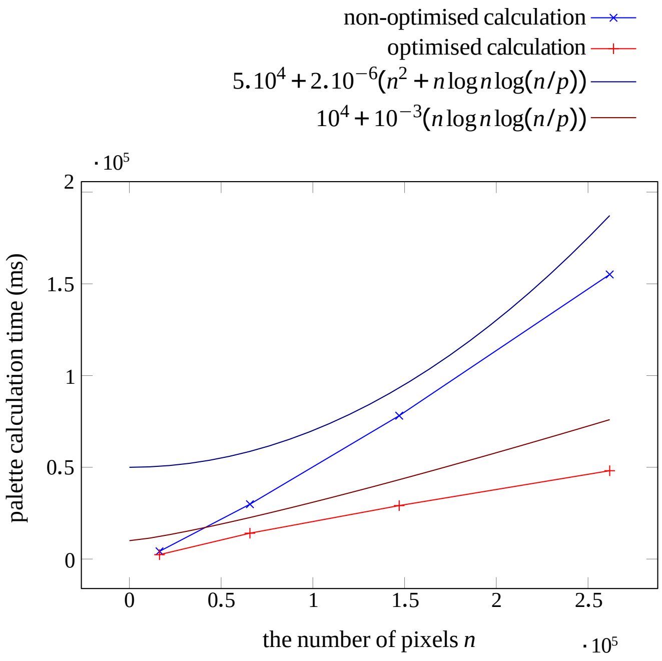


Figure 2: Experimental measurement of the dynamic palette calculation time with and without optimisation, function of n the number of pixels. Theoretical estimations of the worst-case time complexity (with coefficients for visibility) are also plotted for reference.

Experimental evaluation shows that the evolution (slope) of the measured execution times is significantly slower than the theoretical worst-case estimation, which first confirms what has been theoretically established and second is a positive indicator of the performance of the algorithm and of its implementation, and thus of the practicability of the functional paradigm in this case.

4 Application to Dithering

In order to further inspect the practicability of functional programming for image processing, we next consider a dithering (error diffusion) algorithm for images based on a colour palette.

We have selected the well-known Floyd-Steinberg dithering algorithm whose approach is to calculate the difference between the current pixel's colour and the nearest colour inside the palette, and to next diffuse with predefined coefficients this error to neighbours of the current pixel, precisely to the east pixel, south west pixel, south pixel and south east pixel of the current pixel. This algorithm can be implemented in accordance to the functional paradigm with a function that takes as parameters the image pixels as a list of colours, the palette to apply and the image width and height. The returned value is the new image pixels, as a list of colours. Dithering is applied in one single pass, with thus a worst-case time complexity of $O(n)$.

Application of this error diffusion algorithm implementation to a sample image is illustrated in Figure 3.

We have conducted an empirical evaluation in similar conditions as for the experiment of Section 3.3, this time with the photograph of Figure 3 before applying a palette. The following distinct resolutions (in pixels) were used: 591×787 (i.e. $n = 465117$), 443×590 (i.e. $n = 261370$), 296×394 (i.e. $n = 116624$) and 148×197 (i.e. $n = 29156$). The number of colours k was 42 216, 33 210, 23 636 and 9 629, respectively. The palette size p remained fixed at 16. The time taken by the dithering process is shown in Figure 4.

As in the previous experiment, the empirical results show the efficiency of our implementation as the evolution of the measured dithering times is slower than the theoretical estimation.

5 Going Further: Parallel Processing

We complete this constructive proof of the practicability of functional programming for image processing by considering parallel processing.

5.1 Threads

First, we have relied on Racket threads to conduct dithering in parallel. The idea to enable parallel processing for the Floyd-Steinberg dithering algorithm is to divide the image into several consecutive areas and to process each of those in a separate thread. Each thread applies dithering on its area and returns the result. Results are then merged back into one single image.

We have used two threads, in addition to the control (main) thread, for our experiments, with thus the original image divided into what we call the upper half and the lower half. One can note that the first pixel row of the lower half does not fully aggregate error since the previous pixel row, that is the last pixel row of the upper half, is treated separately in another thread, and without resource sharing. Therefore, error is not diffused from the last pixel row of the upper half to the first pixel row of the lower half. It is however merely a remark since this does not produce artefacts and thus goes unnoticed.

An excerpt of our implementation with threads is given in Listings 2 and 3.

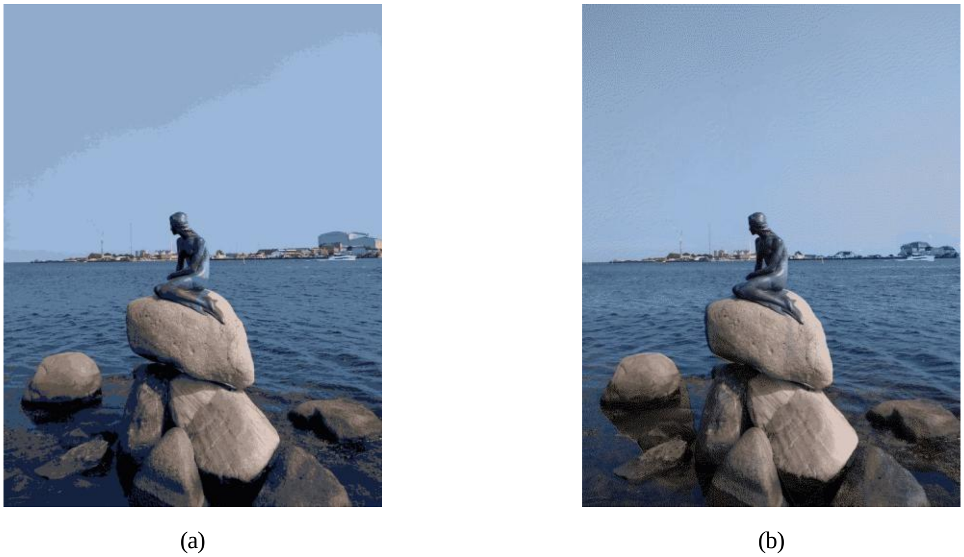
Listing 2: Parallel processing for dithering with threads: thread creation and result reporting to the main thread.

```
(define (create-thread parent-thread half-id half-image palette bitmap-width
        half-bitmap-height)
  (thread (lambda () (let ([half-result (apply-palette-dithering half-image
              palette bitmap-width half-bitmap-height)])
     ; completed: report the result to the main thread, together with
          the half identifier
     (thread-send parent-thread (cons half-id half-result))))))
```

(The function `apply-palette-dithering` applies the dithering algorithm as described in Section 4.)

Listing 3: Parallel processing for dithering with threads: the control (main) thread.

```
(define (dith-thread image-colours palette bitmap-width bitmap-height)
  (let* ([upper-half-height (floor (/ bitmap-height 2))]
```



(a)

(b)

Figure 3: 16-colour palette: (a) no dithering; (b) dithering applied. (Photograph taken by the author.)

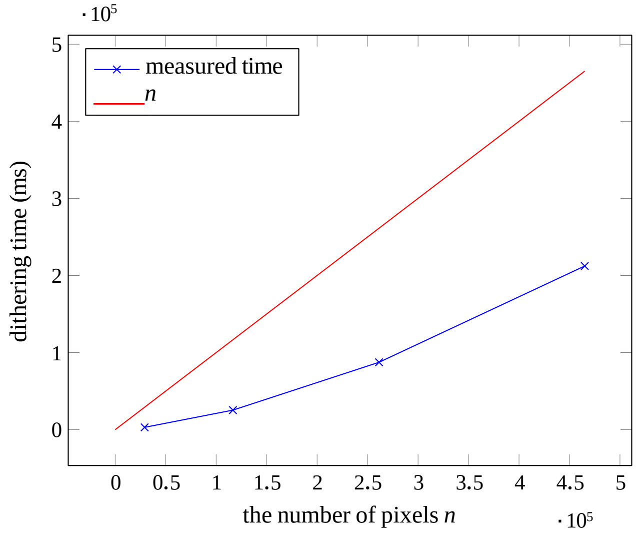


Figure 4: Experimental measurement of the dithering time, function of n the number of pixels. The theoretical estimation of the worst-case time complexity is also plotted for reference.

```
       [lower-half-height (- bitmap-height upper-half-height)])
  (let-values ([(upper-half lower-half) (split-at image-colours (*
      bitmap-width upper-half-height))])
    (let* ([thread1 (create-thread (current-thread) 'upper-half upper-half
          palette bitmap-width upper-half-height)]
           [thread2 (create-thread (current-thread) 'lower-half lower-half
              palette bitmap-width lower-half-height)]
           [id-data1 (thread-receive)] ; receive one half of the result
           [id1 (car id-data1)] [data1 (cdr id-data1)]
           [data2 (cdr (thread-receive))]) ; receive the other half
      (thread-wait thread1) ; for safety as 'thread-receive' signals...
      (thread-wait thread2) ; ... that the thread is about to terminate
      (if (eq? id1 'upper-half) ; merge the results based on the half
          identifier
        (append data1 data2) (append data2 data1))))))
```

Moreover, it is recalled that Racket threads run on one single core of the processor, even if several are available. One should note that our implementation is elegant, short and using thread mail boxes (thread-send, thread-receive). There are no shared resources across threads (no concurrent access), which will be even more important for the next section on futures.

We have conducted an empirical evaluation in similar conditions as for the experiment of Section 4, notably with the same four image files, but this time with the multithreaded implementation. As explained, two threads in addition to the control (main) thread were used. The measured dithering times are summarised in Table 1 together with the experimental results of the sequential implementation for comparison.

Table 1: Experimental measurement of the dithering time induced by the multithreaded implementation, function of n the number of pixels. The results in the case of the sequential implementation are included for reference.

Image resolution (pixels)	Sequential implementation (ms)	Multithreaded implementation (ms)	Speed-up factor
148×197	3 032	2 369	1.28
296×394	25 365	18 041	1.41
443×590	87 420	61 370	1.42
591×787	212 302	148 652	1.43

The empirical results show significant speed-up compared to the sequential implementation. Furthermore, the speed-up value is rather stable at approximately 1.4. Which is remarkable in that

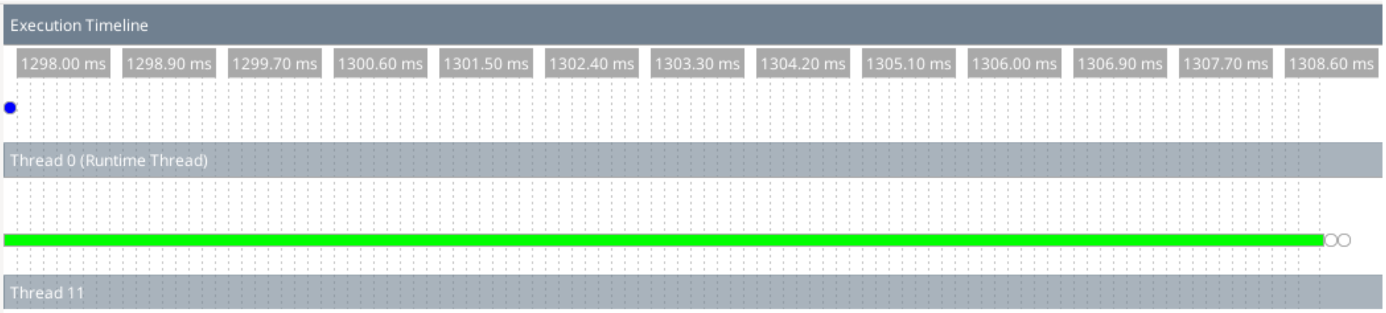


Figure 5: Output of the future visualizer tool: the green bar from start to end of the dithering algorithm execution shows that the main thread and the future's thread successfully run fully in parallel on distinct CPU cores.

as explained, all the Racket threads run on one single processor core.

5.2 Futures

In order to achieve parallel processing, unlike threads, on *several* cores of the processor, Racket provides the "future" mechanism. Parallelism with futures can however become rapidly hampered when information from the main thread is required, thus blocking parallel processing. This is the case, for instance, when I/O operations are conducted.

As with the single core multithreaded approach above, we divide the image into two parts, the upper and lower half. The upper half is treated in parallel by a future, and the lower half in the main thread. Both results are eventually merged and returned. Refer to Listing 4.

Listing 4: Parallel processing on several cores for dithering with futures.

```
(define (dith-future image-colours palette bitmap-width bitmap-height)
  (let* ([upper-half-height (floor (/ bitmap-height 2))]
         [lower-half-height (- bitmap-height upper-half-height)])
    (let-values ([(upper-half lower-half) (split-at image-colours (*
           bitmap-width upper-half-height))])
      (let* ([future1 (future (lambda () (apply-palette-dithering upper-half
               palette bitmap-width upper-half-height)))]
             [lower-half-result (apply-palette-dithering lower-half palette
                    bitmap-width lower-half-height)]
             [upper-half-result (touch future1)])
        (append upper-half-result lower-half-result)))))
```

The results obtained from this multithreaded implementation based on futures show that parallelism on several cores of the CPU has been successfully achieved. When applied to the smallest of the four sample images of the previous experiment, the future visualizer tool output is as shown in Figure 5. The topmost row represents the main thread, for us processing the lower half of the image, and the second row corresponds to the future's thread: it displays a green bar spanning the whole dithering execution time. This uninterrupted green bar means that the corresponding future has been successfully run fully in parallel to the main thread, on a distinct CPU core. This desirable situation is enabled by the absence of shared resources, and communication in general, between the main thread (processing the lower half of the image) and the future's thread (processing the upper half of the image).

We have conducted an empirical evaluation in similar conditions as for the experiment of Section 4, notably with the same four image files, but this time with the multithreaded implementation based on futures. As explained, one future in addition to the control (main) thread was used. The measured dithering times are plotted in Figure 6 together with the experimental results of the sequential and single core multithreaded implementations for comparison.

This empirical evaluation shows that parallel processing on several cores, when successfully achieved by futures as explained (see Figure 5), further significantly reduces the time required to apply the dithering algorithm to the image: as shown in this figure, we measured a 1.87, 1.90, 1.98 and 1.94 speed-up factor for the four images, respectively, compared with the single core multithreaded approach.

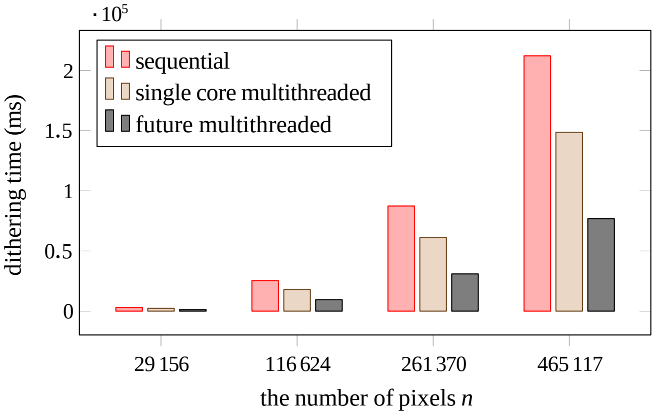


Figure 6: Experimental measurement of the dithering time induced by the implementation based on futures, function of n the number of pixels. The results in the case of the sequential and single core multithreaded implementations are also displayed for reference.

6 Concluding Remarks

Through this work, we have successfully shown that functional programming, with some minor exceptions to the functional paradigm, is capable for image processing. Image processing algorithms, such as dynamic palette calculation and dithering (error diffusion) can be elegantly implemented. Moreover, even parallel processing, with threads on one single core and with futures on distinct physical cores, is feasible and brings significant performance improvements, as quantitatively shown by our experiments.

Besides, parallelization is notoriously challenging for programmers, and often harmful to program robustness. Thanks to the functional paradigm, robustness is retained when implementing parallel computation tasks. Overall, because Racket is a high-level language and tolerates exceptional imperative programming constructs, it features a very high usability, notably confirmed throughout our experiments.

Although we have been able to show to some degree the applicability and practicability of functional programming, and more generally of the functional paradigm, to image processing, the experimentally measured processing times remain relatively high. So, as future work, it would be interesting to next compare the achieved performances and those obtained from an implementation based on an imperative or object-oriented programming language. Such a discussion could also be extended to a pure functional language such as Haskell: it would certainly be more difficult to manipulate, but the existing libraries, like GUI ones for Haskell, could perhaps facilitate implementation. Finally, showing whether graphical animation is possible too, like for interactive content, is yet another interesting subject, with however even less tolerance for long processing times.

References

[1] Antoine Bossard. The SOF programming paradigm: A sequence of pure functions. *International Journal of Software Innovation*, 10(1):1–14, 2022.

[2] Antoine Bossard and Keiichi Kaneko. A new methodology for a functional and logic programming course: On smoothening the transition between the two paradigms. In *Proceedings of the 20th Annual SIG Conference on Information Technology Education (SIGITE; Tacoma, WA, USA, 2–5 October)*, pages 63–68. Association for Computing Machinery, 2019.

[3] Matthias Felleisen, Robert Bruce Findler, Matthew Flatt, Shriram Krishnamurthi, Eli Barzilay, Jay McCarthy, and Sam Tobin-Hochstadt. A programmable programming language. *Communications of the ACM*, 61(3):62–71, February 2018.

[4] Jason Gregory. *Game Engine Architecture*. Taylor and Francis, Boca Raton, FL, USA, 2009.

[5] Till Haenisch. A case study on using functional programming for Internet of Things applications. *Athens Journal of Technology & Engineering*, 3(1):29–38, March 2016.

[6] John Hughes. Why functional programming matters. *The Computer Journal*, 32(2):98–107, 1989.

[7] ISO/IEC JTC 1/SC 22. Programming languages – C++. Technical Report ISO/IEC 14882:2017, 5th edition, International Organization for Standardization, December 2017.

[8] Dimitris Kyriakoudis and Chris Kiefer. uSEQ: A LISPy modular sequencer for Eurorack with a livecodable microcontroller. In *Proceedings of 7th International Conference on Live Coding (ICLC; Utrecht, The Netherlands, 19–23 April)*, pages 1–15. Zenodo, April 2023.

[9] Simon Marlow. *Haskell 2010 Language Report*, April 2010.

[10] Rinus Plasmeijer and Marko van Eekelen. Keep it clean: a unique approach to functional programming. *ACM SIGPLAN Notices*, 34(6):23–31, June 1999.

[11] Franco Raimondi, Giuseppe Primiero, Kelly Androutsopoulos, Nikos Gorogiannis, Martin J. Loomes, Michael Margolis, Puja Varsani, Nick Weldin, and Alex Zivanovic. A Racket-based robot to teach first-year computer science. In Kent M. Pitman, editor, *Proceedings of the 7th European Lisp Symposium (ELS; Paris, France, 5–6 May)*, pages 54–62, 2014.

[12] Jocelyn Se´rot, Georges Que´not, and Bertrand Zavidovique. Functional programming on a dataflow architecture: Applications in real-time image processing. *Machine Vision and Applications*, 7:44–56, December 1993.

[13] Brian Terlson. ECMAScript 2018 language specification. Technical Report ECMA-262, 9th edition, Ecma International, June 2018.

[14] Mikus Vanags and Rudite Cevere. The perfect lambda syntax. *Baltic Journal of Modern Computing*, 6(1):13–30, 2018.

Antoine Bossard is a Professor of the Graduate School of Science of Kanagawa University in Japan. He received the BS and MS degrees from Universite´ de Caen Basse-Normandie, France in 2005 and 2007, respectively, and the Ph.D. degree from Tokyo University of Agriculture and Technology, Japan in 2011. Amongst others, he is in charge of the computer architecture and functional programming lectures for undergraduate students, and of a graph theory lecture for master students. His research activities are focused on interconnection networks (e.g. network topologies, routing problems, fault tolerance) and information representation and processing of Chinese characters (e.g. fingerprinting). He is a Senior Member of ACM and a member of TUG.

Energy-Efficient Dynamic Cluster Formation for WSN Lifetime Optimization

Chaima BENSAID*
Khemis Meliana University, ALGERIA
Mohammed Khalil HADJ AHMED †
Khemis Meliana University, ALGERIA
Mohamed Mehdi BENALI‡
Khemis Meliana University, ALGERIA

Abstract

Wireless sensor networks (WSNs) have become increasingly popular over the last few decades, particularly in environmental monitoring, industrial control, healthcare, security, and offer the possibility of collecting data on physical and environmental phenomena, enabling advances in safety and intelligent surveillance. Each sensor node, equipped with physical sensors, processing circuitry, and wireless communication modules, acts as an autonomous agent capable of monitoring its surroundings and relaying information. However, large-scale deployment of these networks introduces challenges such as efficient energy management, communication reliability, fault tolerance, data security, and dynamic topology management. This article addresses these challenges and proposes innovative solutions to enhance WSN performance and sustainability, focusing on optimizing routing protocols. We propose a new solution to optimizing the AOMDV routing protocol, aiming to improve Quality of Service (QoS) while optimizing energy efficiency and the network lifetime. Additionally, we introduce a new clustering algorithm that dynamically forms clusters based on node density and energy levels to minimize communication overhead and prolong network lifetime. Extensive simulations using the NS2 network simulator demonstrate the effectiveness of our methods in improving QoS, energy efficiency, and network lifetime compared to existing protocols. Our contributions offer promising solutions for enhancing WSN performance and sustainability, enabling broader adoption in diverse applications.

Key Words:WSN; AOMDV; QoS; NS2; Lifetime.

1 Introduction

Wireless sensor networks (WSN) are generally made up of small sensors distributed in a more or less random geographical area known as the catchment area or area of interest. These networks (1) allow data to be collected on physical and environmental phenomena, paving the way for numerous innovations. WSNs are distinguished by their ability to collect, process and transmit data autonomously and collaboratively. Each sensor node (2) equipped with physical sensors, processing circuitry, and wireless communication modules acts as an autonomous agent monitoring its environment and transmitting information to other nodes or a central base station. However, their massive deployment presents technical and engineering challenges, including efficient energy management, communications reliability, fault tolerance, data security, and dynamic topology management.

Routing (3; 5) in such networks is pivotal in ensuring efficient data delivery while optimizing energy consumption and network lifetime. In WSNs, traditional routing techniques encounter difficulties like dynamic network topologies, high energy consumption, and constrained bandwidth. The ability of the Ad hoc On-Demand Multipath Distance Vector (AOMDV) protocol to create several routes between a source and a destination has made it stand out among these protocols.However, we have turned our attention to clustering techniques for further efficiency gains.

Clustering (7) involves organizing sensor nodes into groups or clusters. Where there may be a designated cluster head for each cluster who is in charge of arranging communication both inside and between clusters. The integration of clustering with routing protocols such as AOMDV presents a promising avenue to address the unique challenges of WSNs. By leveraging clustering, network resources can be allocated more efficiently, reducing overhead and prolonging network lifetime. Additionally, clustering facilitates localized data processing and aggregation, mitigating the impact of bandwidth constraints, and enhancing scalability. The present article discusses the synergy between clustering and the AOMDV protocol in WSNs. We explore how clustering enhances AOMDV performance by reducing routing overhead, improving network scalability, and increasing resilience to node failures.

In this article, we explore these challenges and propose innovative solutions to improve the performance and sustainability of WSNs. In particular, we focus on optimizing routing protocols, which are crucial for efficient communication

[*] Computer Science Department.Email: chaima.bensaid@univ-dbkm.dz

[†] Computer Science Department.Email: mi19.m.hadj-ahmed@univ-dbkm.dz

[‡] Computer Science Department.Email: mohamed.benali7@univ-dbkm.dz

ISCA Copyright© 2025

in these networks, often deployed in hostile environments. This Introduction lays the foundations of our research. Section 2 analyses the AOMDV routing protocol and the principle of clustering routing protocols in detail. Section 3 presents related work, while Section 4 describes the implementation and simulation of our clustering algorithm under NS2, demonstrating its energy efficiency and communication reliability improvements. In summary, this work aims to significantly improve routing in WSNs by optimizing routing protocols and implementing innovative solutions to these networks' technical and engineering challenges..

2 AOMDV protocol and clustering techniques

2.1 The AOMDV Routing Protocol

Adhoc On Demand Multipath Distance Vector or AOMDV (3) is a multipath reactive routing protocol. Route maintenance and route discovery are its two primary stages. This multipath routing protocol establishes several disjoint routes without routing loops from source to destination. However, it mainly uses the best path in terms of hop count to transfer data. These multiple paths can be used for load balancing or to provide backup routes in the event of failure of the main route being failed. AOMDV 's (4) primary concept is to compute various routes from the traffic source to the destination while avoiding the formation of routing loops. At the start of the procedure, the source sends the route request message RREQ (Route REQuest) to its adjacent nodes. The adjacent nodes receive the RREQ and send an RREQ to their adjacent nodes. This process continues until the destination node receives the route request (see Figure 1).

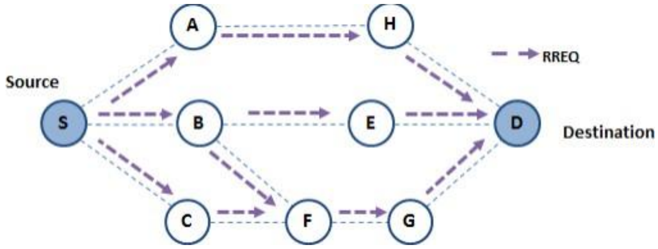
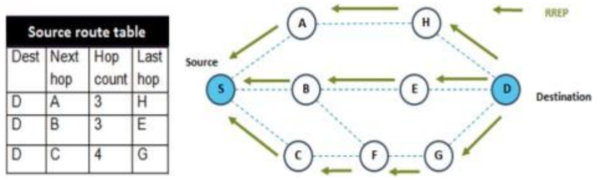


Figure 1: Route Discovery Phase .

The destination (5) node generates an RREP (Route REPly) for each RREQ received. The source node receives several RREPs corresponding to the paths discovered. If only one RREP is received, i.e., only one route is recognized between the source and the destination, then it sends the data packets on this route. Otherwise, if several RREPs (6) have been received, the source chooses the best route, i.e. the one with the lowest hop count. The other routes await the arrival of a RERR packet indicating that the main route has been broken. In this case, the best route among the alternative routes is selected to retransmit the data. If no RREP is received, a new route discovery phase is triggered (see Figure 2)


Source route table

Dest	Next hop	Hop count	Last hop
D	A	3	H
D	B	3	E
D	C	4	G

Figure 2: RERP packet generation

2.2 Clusterig technique in WSNs

The Clustering (7) in networks is a distributed method of dealing with problems such as network lifetime and energy. These problems can be solved by clustering sensor nodes. A cluster head (CH) controls internal communication between sensors in the same cluster. Cluster heads can communicate with each other to reach the sink.

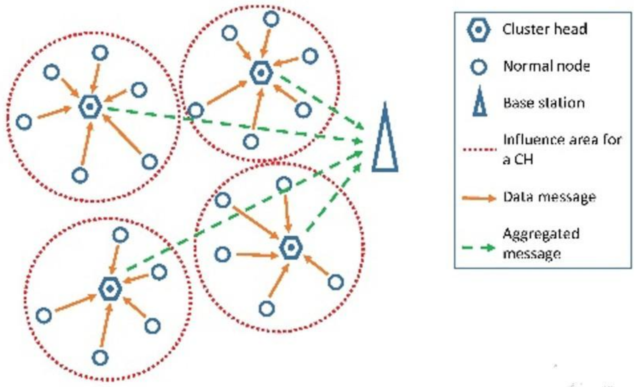


Figure 3: Clustering procedure in a wireless sensor network.

As we can see in Figure 3 are some important components, listed as follows:

- Cluster member (normal node);
- Cluster head;
- Gateway node.

In the wireless network (8), each cluster has a group leader, known as the cluster head (CH), who manages inter-cluster and intra-cluster communication. One of the main metrics of WSNs is the election of cluster heads:

1. large residual energy;
2. large number of neighbors;
3. The distance of nodes from the base station.

Cluster leader nodes (7) link sensor nodes and the base station. The cluster technique in Wireless Sensor Networks (WSNs) offers several advantages, it facilitates data aggregation, which helps reduce data transmission and conserve energy. Additionally, it enhances resource reusability and enables the formation of a virtual backbone for inter-cluster routing through cluster heads and gateway nodes. The cluster (8) structure contributes to a smaller and more stable network, ultimately improving network lifetime and minimizing network traffic. Furthermore, it supports data aggregation and updates at cluster heads while reducing channel contention.

Various clustering (9) routing protocols ely mostly on CH selection schemes . CH selection methods include deterministic, random, and adaptive approaches. Deterministic schemes position CHs at fixed network locations, while random schemes select CHs randomly from sensor nodes. Adaptive schemes base CH selection on location or battery .Clustering mechanisms (8) can be static or dynamic. Static clustering maintains a fixed topology throughout the network's lifetime, leading to quick battery depletion in CHs. Dynamic clustering allows topology changes over time, promoting energy balance across sensor nodes and extending network lifetime.

3 Related Work

Multipath routing protocols (10) rely on multiple paths for packet transmission. One of the best multipath protocols is AOMDV. Although the AOMDV protocol has many advantages, maintaining multiple alternative paths can reduce the battery life of nodes, and may generate more control packets such as error, hold and link discovery messages. We begin with a state-of-the-art review of existing work in the literature.

The authors in (11) sought to balance the battery use of mobile nodes and extend the network lifetime. A new routing protocol called EA-AOMDV is presented.The EA-AOMDV protocol's primary goal is to balance nodal energy consumption to keep one or more crucial nodes from running out of energy and ceasing to be able to communicate with the rest of the network.During the path selection phase, the protocol searches for each route's average residual energy and least nodal energy. The sum of the average and minimum energy is the parameter used to choose the best routes. The outcomes demonstrate enhanced network lifetime, overload, end-to-end latency, and packet delivery ratio performance.

In (12) the authors propose a multi-path reactive routing protocol to save network energy and bandwidth. The optimal routes regarding available bandwidth and the least amount of leftover energy form the basis of the suggested routing protocol. This protocol is incredibly efficient, using less energy and losing fewer packets. However, the battery of nodes on these channels might quickly run out if you constantly rely on the same paths with high bandwidth and low energy.

The authors of (13) , used an optimization algorithm known as particle swarm optimization (PSO) to suggest an AOMDV routing protocol optimized for energy consumption. To cut down on consumption, the system determines the route with the best distance. The energy level is highest on the chosen main path. Upon receiving a route response packet, an intermediate node first determines how much energy is left and adds it to the energy field of the response packet. The sender will select the path with the greatest average energy value if it receives several responses. Simulation results indicate better communication throughput, latency, and node lifetime performance.

In (14), The authors suggest utilizing the multipath routing protocol AOMDV to increase network longevity and energy efficiency. The proposed AOMDV EE protocol uses energy thresholds to choose energy-efficient routes from those available during protocol implementation. The results show that the suggested EE AOMDV protocol is more energy-efficient than the AOMDV protocol. The analysis utilizes network lifetime and energy usage by changing node speed.

A multipath routing technique is suggested by the authors in (15). When generating numerous disjoint pathways, the proposed protocol, known as AOMDV-FF, considers the distance between the source and the destination and the residual energy of nodes. According to simulation results, AOMDV-FF has superior throughput and overhead.

The LEACH protocol is improved by LEACH-VD (22).to lower energy usage. Three steps make up the protocol's operation. The initial stage is to create clusters and choose cluster heads using the LEACH protocol. The shortest routes between each cluster head are then identified. Lastly, the shortest pathways are found using the DIJKSTRA algorithm. The energy consumption of the DIJKSTRA algorithm, which determines the shortest pathways, is one of LEACH-VD's primary disadvantages.

TEEN-V (23).enhances the TEEN protocol to reduce data transmission rates. It works in two stages: firstly, running the TEEN protocol to establish clusters and designate cluster leaders, and secondly, calculating the shortest paths between cluster leaders to save energy. However, a notable drawback of TEEN-V is the energy consumption associated with the vector quantization process.

The authors of (25) uses Dijkstra's algorithm to optimize energy in WSNs. Dijkstra's algorithm is adapted to minimize energy consumption by selecting routes that balance the energy load between nodes. This work focuses on how graph-based algorithms can improve energy utilization in WSNs, providing a promising approach for long-lived, sustainable sensor network deployments.

4 The proposed protocol

There are several particular difficulties in creating an effective routing system for wireless sensor networks (WSNs)(16). Power management is one of the main problems. In a WSN, routing is essential, enabling individual nodes to transmit data captured from sources to destinations via intermediate nodes. Each transmission consumes energy, from data or as a relay for other nodes. Consequently, a routing algorithm must incorporate energy management mechanisms to minimize node consumption and extend the network lifetime.Furthermore, an effective routing system needs to consider the unique properties of sensors, such as energy resource limitations and hardware constraints. developing routing strategies that adapt to these constraints is crucial to meet application requirements and optimize network performance.

This section describes the simulation and implementation of our clustering algorithm in WSNs using NS2. We explain the steps and tools adopted to implement and evaluate our approach, which aims to optimize energy consumption and increase network lifetime. In our solution, proposed Sensor networks comprise sensors. During simulations, sensors rapidly exhaust energy. Sensors are randomly distributed over a 1000 m x 1000 m area. The base station is situated at the central location of the network. The network's sensors are also uniform. Moreover, the sensors have a data to send to the base station using a communication file with a simple energy model. The protocol consists of two main stages(see Figure 4): configuration and communication.

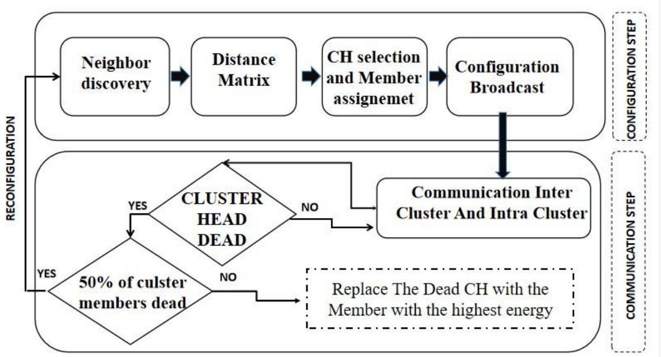

Neighbor discovery
Distance Matrix
CH selection and Member assignemet
Configuration Broadcast
CONFIGURATION STEP
RECONFIGURATION
CLUSTER HEAD DEAD
YES
NO
Communication Inter Cluster And Intra Cluster
YES
50% of culster members dead
NO
Replace The Dead CH with the Member with the highest energy
COMMUNICATION STEP

Figure 4: Proposed protocol architecture

Enhancing energy conservation, a crucial factor in prolonging the life of a sensor network, is the goal of our solution. Clusters are formed and their cluster leaders are chosen using a hybrid approach. formation of clusters using a novel centralized algorithm. Based on the distances between sensors and a predetermined threshold, our algorithm clusters the closest sensors.

There are two primary steps in our algorithm:
- A centralized network configuration stage, which is repeated multiple times over the network's lifetime.
- A distributed communication step between the base station and the network sensors.

In the first stage, the BS divides the network into clusters using the Euclidean distance between sensors. This division is based on an adaptive and dynamic algorithm, enabling efficient allocation of network resources.The second stage, communication, establishes communications between the network's sensors and the base station, based on the clusters formed in the first stage. This hybrid approach aims to optimize energy utilization while ensuring efficient network coverage by all nodes. In the following sections, we will detail each step of the algorithm and explain in detail how it works.

4.1 Configuration stage

The configuration stage is carried out in offline mode before communication.. In this stage, the base station divides the network into several clusters according to 4 phases: from neighbor detection to the assignment of member nodes to their group leader (CH).

4.1.1 Neighbor discovery phase

In this phase, nodes broadcast a HELLO message to discover adjacent nodes in a single hop. The result of this phase is to generate an adjacency table for each node. The aim of this step is to generate the 1-hop neighbors from which the RREQ packet will be broadcast. Once all nodes generate their neighbor table the configuration step starts to generate the cluster heads.

4.1.2 Distance matrix creation phase

In this phase, the base station creates a distance matrix between nodes based on Euclidean distance. We can represent the Euclidean distance matrix D as an $m \times m$ matrix where the element D_{ij} represents the distance between node i and node j. Thus :

$$D = \begin{pmatrix} d(\mathbf{p}_1, \mathbf{p}_1) & d(\mathbf{p}_1, \mathbf{p}_2) & \cdots & d(\mathbf{p}_1, \mathbf{p}_m) \\ d(\mathbf{p}_2, \mathbf{p}_1) & d(\mathbf{p}_2, \mathbf{p}_2) & \cdots & d(\mathbf{p}_2, \mathbf{p}_m) \\ . & . & \ddots & . \\ d(\mathbf{p}_m, \mathbf{p}_1) & d(\mathbf{p}_m, \mathbf{p}_2) & \cdots & d(\mathbf{p}_m, \mathbf{p}_m) \end{pmatrix}$$

where each element of the matrix is calculated as follows :

$$D_{ij} = d(\mathbf{p}_i, \mathbf{p}_j) = \sqrt{\sum_{k=1}^{n} (p_{ik} - p_{jk})^2}$$

4.1.3 CH selection and Member assignemet

In this phase, the base station selects the group leaders based on the distance matrix and the distance threshold K according to our CH selection algorithm.The distance threshold K is a value used to determine whether a node can be selected as a group leader. This value represents the minimum distance between two clusters, and our algorithm generates a list of cluster heads whose distance between each two cluster heads must exceed or equal the value of K.

The following algorithm represents the CH selection technique

Algorithm 1 Node clustering

```
Determine the distance matrix’s maximum value.
Store the distance matrix’s greatest value’s indices i and j.
Initialize CHlist with i and j {First cluster heads}
for Each node in the network do
    boolean isClusterHead ← true
    for each cluster head in CH do
        if distance between node and cluster head ≤ K then
            isClusterHead ← false
            break
        end if
    end for
    if isClusterHead then
        Add node to CH list as new cluster head
    end if
end for
for  each node from 0 to n do
    Find the nearest head cluster using the distance matrix
    Assign element to clustet head
end for
```

Now that we have all the CHs, nodes will be assigned to their nearest cluster based on Euclidean distance. A configuration message containing the node identification and its cluster will be sent.

4.1.4 Configuration Broadcast phase

The broadcasting of a configuration message with the sensors’ identity completes the configuration process. Each sensor only keeps the data it needs due to issues with sensor memory limitations. Each sensor retains only the data pertaining to the cluster in which it is situated. The cluster head keeps the list of nearby cluster heads. During the communication step, each cluster head takes charge of communication between the clusters.

4.2 Communication stage

Two layers of communication exist in a WSN: inter-cluster communication between cluster heads and the base station and intra-cluster communication between clustrer members and cluster leaders. At this point, nodes in a cluster exchange information with their cluster leader to send data to the base station. The cluster manager (CH) manages communication between the clusters and relays data between them every time period. The base station re-evaluates sensor status, such as power consumption. If the CH is destroyed, we call this a reconfiguration phase.Algorithm 2 presents the reconfiguration technique

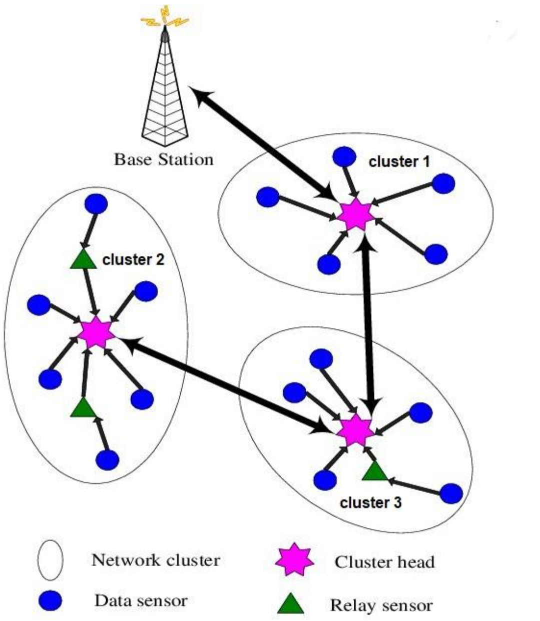


Figure 5: Clustered communication architecture for a sensor network

Algorithm 2 Cluster reconfiguration

```
TH = Threshold for rejected clusters (minuscules) 5%
X = Percentage of living cluster members
for each CH in CHlist  do
    if $Energy_{CH} \leq TH$ & $X \geq 50\%$ then
        Identify the node with the highest energy.
        Designate this node as the new cluster leader.
        Update cluster assignments accordingly.
    end if
    if $Energy_{CH} \leq TH$ & $X \leq 50\%$ then
        Eiminating dead nodes
        Repeat configuration phase
    end if
end for
```

This procedure identifies the node with the highest energy within the low-energy CH cluster and designates it as the new cluster leader. It then updates the cluster assignments accordingly.

5 Results and Interpretation

This section presents the results of simulations performed on the AOMDV protocol without our approach, and with our approach on WSNs. We begin by defining the various simulation parameters, followed by an interpretation and discussion of the results obtained.

5.1 Simulation parameters

Wireless sensor networks count 140 sensors. The sensors are randomly distributed over 1000 × 1000 meters, with the base station in the center. All sensors are homogeneous, monitoring the environment and continuously sending data to the base station. Data processing power is neglected, excepte for data aggregation at thecluster-head level. The simulation environment is NS2.35. The AOMDV routing protocol uses four main functions: recvReq, recvRep, sendReq, and sendRep. It should be noted that the sendReq function consumes twice as much energy as the other functions. Simulations were carried out in the NS2 simulator. They were run on a test set comprising a sample of 140 randomly generated sensors. The basic parameters of the simulations are shown in the following table:

Parameters	Value
Catchment area	1000 × 1000
BS Position	500 × 500
Number of sensors	140
Number of communication	30
Packet size	512 bits
simulation time	180s
Threshold K	150, 200, 250, 300
Pause Time s	30, 90,120, 180
Initial energy	2500 (μJ)
Threshold for rejected clusters	5%
E_elec	50 nJ/bit
E_amp	1 pJ/bit/m^2
RREQ size	240 bits
RREP size	240 bits
SEND_REQ size	400 bits
END_REP size	400 bits
Collecting radius	20.0 m

Table 1: simulation parameters

5.2 Network performance

Metrics are routing protocol test parameters that allow us to measure the performance of a routing protocol, from which comparisons between protocols can be made. In our study, we considered the following metrics:

5.2.1 Packet delivery ratio

This parameter (17) represents the percentage of packets delivered to their destinations in relation to the number of packets sent into the network.

5.2.2 Average data packet latency

This is the average time required successfully deliver data packets from source to destination, including latency in including latency in queues, buffer storage time (18)

5.2.3 Dropped packets

This is the number of packets ignored or dropped by the CSF (18) .

5.2.4 The network lifetime :

is defined (19) as the time elapsed until all cluster heads (CH) have died. If tdeath CH is the death time of the last cluster leader, then the network lifetime Tli f e is given by:

$$T_{life} = t_{death\_CH}$$

where t_{death_CH} is measured in seconds (s).

5.2.5 Energy consumption

In the AOMDV protocol used WSNs, the energy consumption for different operations such as route request (RREQ), route response (RREP), request sending and response sending can be estimated as a function of various factors such as the energy consumption characteristics of sensor nodes and the length of messages sent (20).

To provide a detailed response, we need to consider :

1. Energy model:Commonly used models include the first-order radio model and the two-beam ground model. For simplicity, let’s assume the first-order radio model first-order(21);
2. Message size:The size of RREQ, RREP and data messages (send request/send response) in bytes.;
3. Transmission and reception energy: Transmission energy (E_{tx}) and reception energy (E_{rx}) can be calculated using the following formulas (21):

$$E_{tx}(k, d) = E_{elec} \times k + E_{amp} \times k \times d^2$$

$$E_{rx}(k) = E_{elec} \times k$$

where E_{elec} is the power consumption per bit to operate the transmitter or receiver circuit, E_{amp} is the power consumption per bit per square meter for the transmission amplifier, k is the message size in bits, and d is the distance between transmitter and receiver (pickup radius) .

4. Route request (RREQ)

Transmission energy for RREQ ($E_{tx,RREQ}$) :

$$E_{tx,RREQ} = E_{elec} \times k + E_{amp} \times k \times d^2$$

Receiving energy for RREQ ($E_{rx,RREQ}$) :

$$E_{rx,RREQ} = E_{elec} \times k$$

5. Route response (RREP)

Transmission energy for RREP ($E_{tx,RREP}$) :

$$E_{tx,RREP} = E_{elec} \times k + E_{amp} \times k \times d^2$$

Receiving energy for RREP ($E_{rx,RREP}$) :

$$E_{rx,RREP} = E_{elec} \times k$$

6. Send request (SENDREQ)

Transmission energy for SENDREQ ($E_{tx,SENDREQ}$) :

$$E_{tx,SENDREQ} = E_{elec} \times k + E_{amp} \times k \times d^2$$

Receiving energy for SENDREQ ($E_{rx,SENDREQ}$) :

$$E_{rx,SENDREQ} = E_{elec} \times k$$

7. Sending response (SENDREP)

Transmission energy for SENDREP ($E_{tx,SENDREP}$) :

$$E_{tx,SENDREP} = E_{elec} \times k + E_{amp} \times k \times d^2$$

Receiving energy for SENDREP ($E_{rx,SENDREP}$) :

$$E_{rx,SENDREP} = E_{elec} \times k$$

These formulas are used to calculate the energy consumption for transmission and reception operations in the AOMDV protocol, taking into account message size and transmission distance.

5.3 Influence of distance threshold (K)

In this section, we examine the impact of distance thresholds on several key metrics in wireless sensor networks. The distance threshold is crucial in many clustering and routing protocols, determining the communication range between nodes and clusters formed. We focus in particular on its influence on energy and sensor lifetime, and on performance metrics such as Packet Delivery Ratio, number of lost packets, and delay in data transmission. Understanding how the choice of distance threshold affects these metrics is essential for designing sensor networks and optimizing their performance in various application scenarios. We explore these relationships empirically through simulation experiments, offering valuable insights for optimizing wireless sensor networks.

5.3.1 Influence of (K) on Formed Clusters

One of the crucial aspects of wireless sensor network optimization is how the distance threshold (K) influences cluster formation. Indeed, K plays a decisive role in the delimitation of clusters and how sensors group according to their spatial proximity. When K is higher, this implies that the communication range between nodes is greater, which can lead to larger clusters and a reduction in the total number of clusters formed. On the other hand, a lower distance threshold will favor the formation of smaller and more numerous clusters, as only nodes that are very close to each other will be grouped together in the same cluster. Thus, understanding the impact of K on cluster formation is essential for designing network architectures tailored to specific application requirements, and for maximizing communication efficiency and energy management in wireless sensor networks. From the graph, we can see that as the threshold decreases, the umber of clusters formed increases. Next, we'll examine how the threshold can influence energy and other metrics(see figure 6)

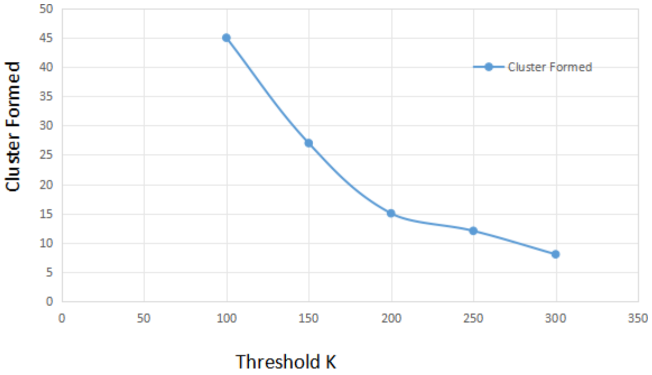


Figure 6: Influence of (K) on Formed Clusters

From the graph, we can see that as the threshold decreases, the number of clusters formed increases. Figures 7, 8 and 9 show clustered data with cluster heads and connections with K=100, 200 and 300, respectively.

Next, we'll look at how the threshold can influence energy and other metrics.

5.3.2 Influence of (K) on the metrics

- Energy consumed : we'll examine how the parameter K influences the energy consumed by the sensors over a 90- second period, as illustrated in figure 10,When K is reduced, more clusters are added, resulting in higher energy consumption. This is because more communication occure between the nodes, requiring higher energy consumption. We note that our approach has improved, as there is a big difference between using clustering and not. In particular, the total energy consumed

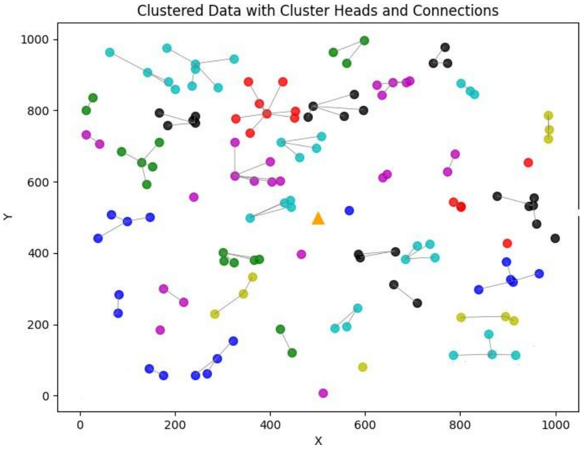


Figure 7: clustering with k = 100

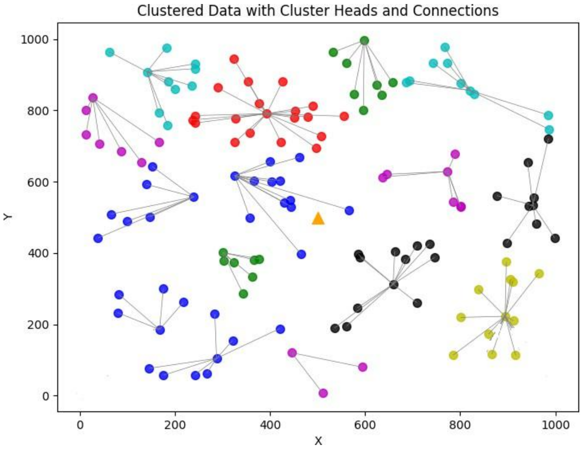


Figure 8: clustering with k = 200

by the sensors is significantly affected by the variation in K.

- PDR and Dropped packets:This section will examine how the threshold influences packet delivery rate (PDR) and packet loss. PDR is a key measure of network performance, representing the percentage of correctly received packets relative to the total number of packets sent (see figure 11,12)

  As we can see, the importance of the threshold parameter, noted k, in the quality of our network is undeniable. Indeed, an increase in k leads to an increase in the number of clusters formed, which intensifies communication between nodes and can influence the number of packets dropped. To illustrate this point, let's consider the following three cases:

  - For k = 300, we observe 4499 dropped packets.;
  - For k = 150, this number rises to 5208 dropped packets;

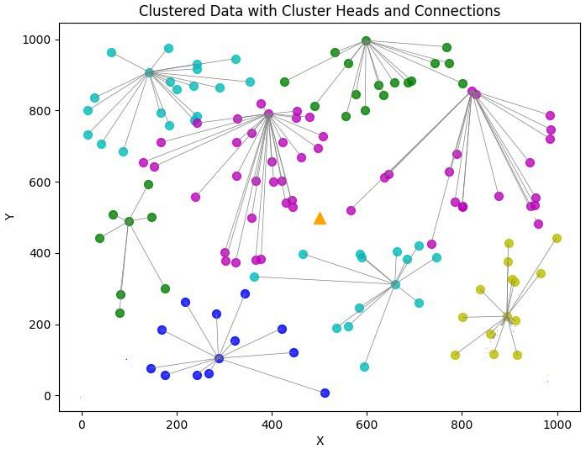


Figure 9: clustering with k = 300

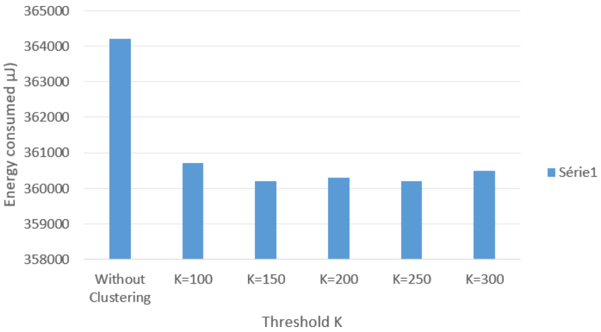


Figure 10: Energy consumed at time 90 seconds per threshold

  It's therefore clear that DROPED150 ¿ DROPED300, which corresponds to 300 ¿ 150. However, the Packet Delivery Ratio (PDR) can also vary as a function of k. For example, AOMDV without clustering, the PDR is 43%, while with a threshold of k = 150, it can increase to 49%. This underlines the crucial importance of choosing our threshold wisely to optimize network quality, as measured by PDR. An inappropriate threshold can degrade network performance by affecting latency, reliability and overall communication efficiency. Therefore, it is crucial to carry out in-depth analysis and rigorous testing to determine the optimum threshold that will guarantee a perfect balance between the number of clusters and the communication load, thus maximizing the performance and quality of our network.

- delay: In figure 13 , We observed that the addition of our contribution increased the delay. For example, when we take the delay for k = 150, it is 8.7ms, whereas without clustering, it is 3.8ms. These two values are not significantly different. This relates to the fact that sometimes the packets from a node are very far from the

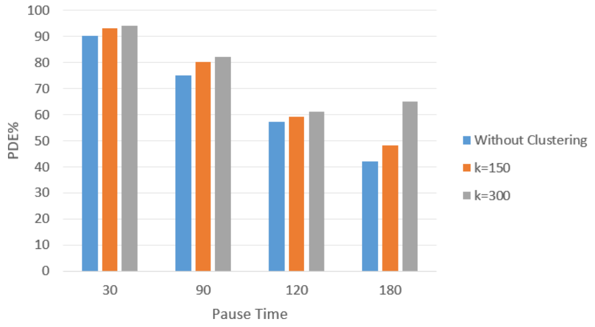


Figure 11: PDR versus pause time

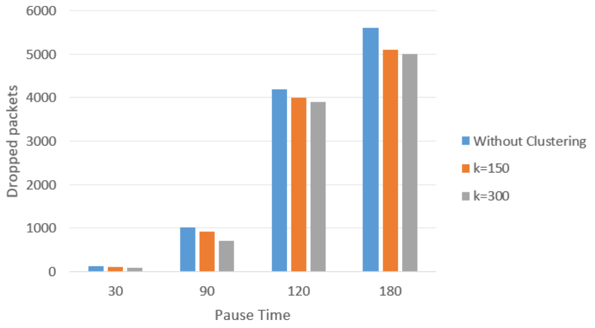


Figure 12: Ignored or dropped packets versus pause time

base station. As the AOMDV algorithm uses the shortest distance without clustering, we notice that packets are delivered quickly, but energy is not considered. However, in our approach, delivery may take a little longer, but with lower energy consumption. It is, therefore, essential to note that the threshold impacts the lead time. A judicious choice of threshold can help balance delay and energy consumption, thus optimizing our network performance.

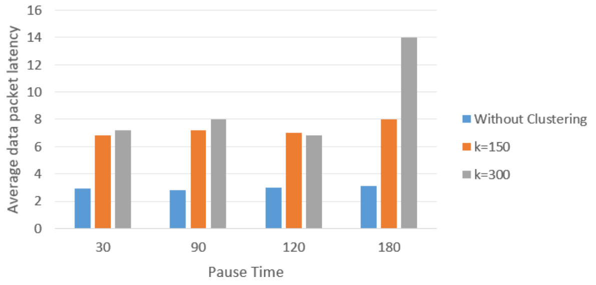


Figure 13: End to End delay versus pause time

5.4 Comparison between Our Approach and AOMDV:

we'll look at how our approach improved total sensor energy consumption and network lifetime. Optimizing the energy consumed by the sensors is a critical criterion for extending the lifetime of the sensors. For this comparison, we are interested in energy consumption per second throughout the 180-second simulation.

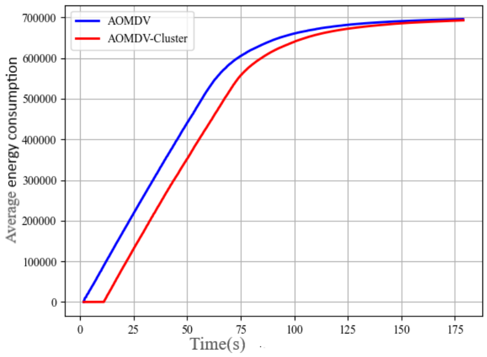


Figure 14: AOMDV VS AOMDV-cluster k = 250

From the figure 14, it's clear that the power consumption of the standard AOMDV quickly reaches a high level. In contrast, our approach shows a more moderate energy consumption and a slower increase. This demonstrates that our method is more efficient in terms of energy management, thus extending the life of the sensors. As a result, it is clear that our approach has improved the total energy consumption of the sensors, automatically leading to an increase in network

The notion of a wireless sensor network's lifetime has several definitions, In our study, we have chosen to define network lifetime in terms of the number of dead sensors at the end of the simulation. Figure 15 illustrates our approach compared with the normal AOMDV protocol, showing the number of dead sensors as a function of time (in seconds).

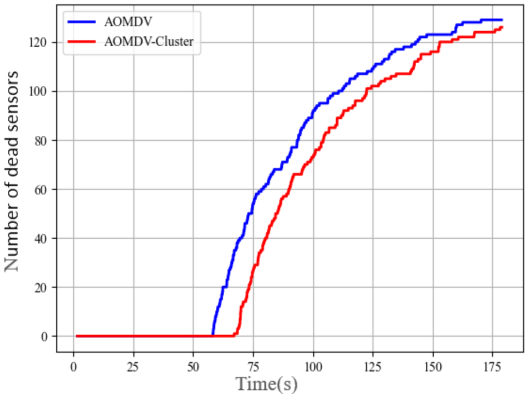


Figure 15: Comparison of the number of dead sensors between our optimized approach and the normal AOMDV protocol as a function of time.

5.5 Comparison with the MCL-BCRP approach:

In this section, we examine a comparison between our approach and that presented in (24). The latter uses the MCL-BCRP (Markov Cluster algorithm - Base Station Controlled Relay Protocol) algorithm for cluster formation, combining the MCL algorithm for this task with a new strategy for selecting sensor nodes acting as Cluster-Heads (CH), based on their location and residual energy level. In addition, the study in (24) compares the MCL-BCRP protocol with several other protocols commonly used for Wireless Sensor Networks (WSNs), including LEACH, TEEN and PEGASIS.

5.5.1 Comparison metrics

In this comparative analysis, we considered the following metrics:

- The number of clusters formed.
- The number of dead clusters at the end of the simulation.

This approach enables us to better assess the performance of our proposal by comparing it with that of an existing protocol such as MCL-BCRP, using objective and meaningful metrics for wireless sensor networks.

5.5.2 Simulation parameters

For this comparison, the simulation parameters used are as follows:

Parameters	Value
Catchment area	1000 × 1000
BS position	500 × 500
Number of sensors	100
Initial energy	$0.5J$
Threshold K	100
Radius of capture	20
Electrical energy(E_{elec})	50×10^{-8} J/bit
Amplification energy(E_{amp})	1.3×10^{-12} J/bit/m^2

Table 2: Simulation parameters

5.5.3 Results obtained

We obtained the following results for our approach (AOMDV-CLUSTER) compared to the MCL-BCRP (24) approach:

Protocol	MCL-BCRP	AOMDV-CLUSTER
Clusters formed	31	12
Dead clusters	14	3
Energy consumed	0,5 J	0,442284 J
Clusters alive at end	55%	75%

Table 3: Comparison results

5.5.4 Comparison

Our results show that our approach (AOMDV-CLUSTER) formed fewer clusters than the MCL-BCRP approach, but with significantly fewer inactive clusters at the end of the simulation. Furthermore, although our approach consumed slightly less energy than MCL-BCRP, it managed to maintain a higher percentage of active clusters until the end of the simulation, indicating greater efficiency in the use of network resources.

6 Conclusion

The reactive AOMDV protocol is efficient and can reduce end-to-end delays, packet loss and increase the rate of successful packet delivery. However, it still suffers from high power consumption due to its flat architecture. This paper uses a preventive approach based on node clustering to reduce energy consumption and increase network lifetime. The proposed AOMDV-clus protocol clusters the network into groups and selects a group leader from each cluster using a new technique. This division allows us to reduce network overload and consequently reduce the number of packets processed by each sensor, thereby increasing the lifetime of a sensor. Our work's simulation results revealed our algorithm's effectiveness in reducing energy consumption and improving the lifetime of WSNs.. More specifically, we observed a significant reduction in energy consumption and an increase in packet delivery rate compared with existing approaches. Indeed, an appropriate distance threshold enables optimal cluster size, minimizing redundant communications and energy consumption. In conclusion, our clustering algorithm proves to be an effective solution for energy management and performance optimization in WSNs. Implementing this approach in NS2 validates the efficiency of our algorithm and highlights its advantages over other existing approaches. Simulation results show a reduced number of dead nodes, a reduction in network overload and energy consumption, anda higher level of performance. and energy consumption, and an improved packet delivery success rate.

7 Limits and future work

Sensors used in WSNs often have limited capabilities in terms of energy storage and management. We could not cover the impact of other types of energy harvesting technologies (such as solar or kinetic powered sensors).

The protocols used for WSNs (e.g., ZigBee, LoRa, NB-IoT) play a key role in energy management. Since the work focuses on the AOMDV protocol, we may not address the energy implications related to other types of protocols. The energy consumption related to data transmission is a key factor that can be influenced by the choice of protocol.

In energy management of sensor networks, security of data and communications is essential. The work does not consider security issues related to energy management (e.g., denial of service attacks affecting sensors), this can be a significant limitation.

For future work in energy management of sensor networks, here are some improvements that could be considered to further this topic:

- Integration of AI and ML technologies for more intelligent and adaptive optimization solutions.
- Optimization of communication between sensors
- Analysis of the impact of varied environments innovative approaches to ensure that energy management systems meet security standards while minimizing the risks of attack

References

[1] Julien-Vergonjanne, A., Sahugue`de, S., Chevalier, L. (2016). Optical Wireless Body Area Networks for Healthcare Applications. In: Uysal, M., Capsoni, C., Ghassemlooy, Z., Boucouvalas, A., Udvary, E. (eds) Optical Wireless Communications. Signals and Communication Technology. Springer, Cham. https://doi.org/10.1007/978-3-319-30201-026

[2] Xu, L., Collier, R., O'Hare, G. M. P. (2017). A Survey of Clustering Techniques in WSNs and Consideration of the Challenges of Applying Such to 5G IoT Scenarios. IEEE Internet of Things Journal, 4(5), 1229–1249. doi:10.1109/JIOT.2017.2726014

[3] Mahak Singla and Paramjeet Singh. Enchancing QoS In MANETs Using Preemptive AOMDV.Global journal of computer science and technology,2018, https://api.semanticscholar.org/CorpusID:196211010

[4] Surabhi Patel, Heman Pathak, "A Cross-Layer Design and Fuzzy Logic based Stability Oriented Routing Protocol", International Journal of Computer Network and Information Security(IJCNIS), Vol.14, No.2, pp.54-66, 2022. DOI: 10.5815/ijcnis.2022.02.05

[5] M. Tekaya, N. Tabbane and S. Tabbane, "Multipath routing mechanism with load balancing in ad hoc network," The 2010 International Conference on Computer Engineering Systems, Cairo, Egypt, 2010, pp. 67-72, doi: 10.1109/ICCES.2010.5674892.

[6] Loo, J., Lloret Mauri, J., Ortiz, J.H. (Eds.). (2012). Mobile Ad Hoc Networks: Current Status and Future Trends (1st ed.). CRC Press. https://doi.org/10.1201/b11447

[7] Sirsikar, S., Wankhede, K. (2015). Comparison of Clustering Algorithmsto Design New Clustering Approach. Procedia Computer Science, 49, 147–154. doi:10.1016/j.procs.2015.04.238

[8] Heinzelman, W. R., Chandrakasan, A., Balakrishnan, H. (2000). Energy-efficient communication protocol for wireless microsensor networks. Proceedings of the 33rd Annual Hawaii International Conference on System Sciences. doi:10.1109/HICSS.2000.926982

[9] Diery Ngom.Lifetime optimization in wireless sensor networks under coverage and network connectivity constraints. PhD thesis, Universite´ de Haute Alsace-Mulhouse; Universite´ Cheikh Anta Diop (Dakar), 2016.

[10] Uttara Korad, Krishna M. Sivalingam.Reliable data delivery in wireless sensor networks using distributed cluster monitoring, International Journal of Sensor Networks 2006 Vol.1 No.1/2

[11] Mahmoud M Shawara, Amany M Sarhan et Nawal A Elfishawy. Energy aware ad-hoc on demand multipath distance vector (EA-AOMDV). In 2017 13th International Computer Engineering Conference (ICENCO), pages 317–322, 2017.

[12] Sivaraman. EE-BWA-AOMDV : Energy Efficient and Bandwidth Aware On-demand Multipath Routing protocol for Mobile Ad hoc Networks. International Journal of Computer Application, vol. 6, no. 2, pages 2250–1797, 2016.

[13] Aqeel Taha, Raed Alsaqour, Mueen Uddin, Maha Abdelhaq et Tanzila Saba. Energy efficient multipath routing protocol for mobile ad-hoc network using the fitness function. IEEE access, vol. 5,pages 10369–10381, 2017.

[14] Soorya V Nair , Shijin Knox G U, 2019, Energy Efficiency and Network Lifetime Improvement in MANET using AOMDV, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH TECHNOLOGY (IJERT) Volume 08, Issue 08 (August 2019),

[15] P Monika, P Venkateswara Rao, BC Premkumar et Pradeep Kumar Mallick. Implementing and Evaluating the Performance Metrics Using Energy Consumption Protocols in MANETs Using Multipath Routing-Fitness Function. In Cognitive Informatics and Soft Computing, pages 281–294. Springer, 2020.

[16] Alazzawi, L., Elkateeb, A., Performance Evaluation of the WSN Routing Protocols Scalability, Journal of Computer Networks and Communications, 2008, 481046, 9 pages, 2008. https://doi.org/10.1155/2008/481046

[17] Mouna Rekik. Routage ge´ographique multi-chemin base´ sur l'intelligence d'essaim pour re´seaux de capteurs et d'actionneurs sans fil : Application aux Smart Grids. Re´seaux et te´le´communications [cs.NI]. Universite´ Lille 1; Universite´ de Sfax, 2016. Franc¸ais. ffNNT : ff. fftel-01370723f

[18] Bensaid, Chaima, et al. "Detection and Ignoring of Blackhole Attack in Vanets Networks." IJCAC vol.6, no.2 2016: pp.1-10. https://doi.org/10.4018/IJCAC.2016040101

[19] Wang Li, Xiangfang. Muiti-layer optimization in wireless ad hoc networks. Retrieved from https://doi.org/doi:10.7282/T3S182W3

[20] Bing Zeng, Yan Dong,An improved harmony search based energy-efficient routing algorithm for wireless sensor networks, Applied Soft

Computing,volume 41,2016,Pages 135-147,ISSN 1568-4946, https://doi.org/10.1016/j.asoc.2015.12.028.

[21] Leila Abbad, Azzedine Nacer, Houda Abbad, Mohammed Taieb Brahim, Nadjet Zioui, A weighted Markov-clustering routing protocol for optimizing energy use in wireless sensor networks, Egyptian Informatics Journal, Volume 23, Issue 3, 2022, Pages 483-497, ISSN 1110-8665, https://doi.org/10.1016/j.eij.2022.05.001.

[22] Mukherjee, P., Pattnaik, P. K., Panda, S. N. (2020). IoT and WSN Applications for Modern Agricultural Advancements: Emerging Research and Opportunities. https://www.igi-global.com/gateway/book/218569

[23] Samant, T., Mukherjee, P., Mukherjee, A., Datta, A. (2017). TEEN-V: A solution for intra-cluster cooperative communication in wirelesssensor network. International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 209–213. doi:10.1109/I-SMAC.2017.8058340

[24] Mohammed Taieb Brahim, Houda Abbad, Sofiane Boukli Hacene: A Low Energy MCL-Based Clustering Routing Protocol for Wireless Sensor Networks. Int. J. Wirel. Networks Broadband Technol. 10(1): 70-95 (2021)

[25] L.Mohammed, A. Mudheher,E.Hamza, Optimizing Energy Efficiency in Wireless Sensor Networks Using Dijkstra's, Algorithm,Instrumentation Mesure Me´trologie, Page: 307-316,DOI: https://doi.org/10.18280/i2m.230406, 2024

Chaima BENSAID is an associate professor at the University of Djilali Bounaama, Khemis Meliana, Algeria. BENSAID obtained her Ph.D. in computer science in February 2020 from Sidi bel Abbes university, Algeria. Her research interests range from on-nerwork security, routing protocol optimization in IoT.

Mohamed Khalil HADJ AHMED received a master's degree in software engineering and distributed systems in 2024 from the computer science department of the university Djilali Bounaama , Khemis Meliana, Algeria. Her research interests focus on energy optimization in WSN, software development nerwork security , Data Science and Machine Learning

Mohamed Mehdi BENALI received a master's degree in software engineering and distributed systems in 2024 from the computer science department of the university Djilali Bounaama , Khemis Meliana, Algeria. Her research interests focus on energy optimization in WSN, software development,Big data, Data Mining , Data Science and Machine Learning

Data analytics enhance decision-making processes effectively using HoQ tool model

Galal eldin Abbas Eltayeb*

Department of Management Information Systems, College of Business and Economics, Qassim University, Buraydah, Saudi Arabia

https://orcid.org/0000-0003-3778-2061

g.eltayeb@qu.edu.sa

Abstract

Data analytics play an important role in promoting effective decision-making by providing valuable insights from data-based analysis. The use of tools such as quality function deployment (QFD) could further simplify this process by translating customer requirements into specific technical requirements, aligning business strategies with customer requirements, and facilitating informed decision-making with data, especially in complex manufacturing systems. In this paper, the proposed model is based on the use of QFD in a way that helps to improve decision-making and provide the ability to describe strategic priorities and express requirements. In this context, we propose developing an application that identifies the organization's requirements starting with the strategy definition to the deployment of the maintenance improvement actions by implementing a house of quality (HoQ) model. In the building of our application, we take into account the use of advanced programming languages through several phases. Initially, the appropriate proposed model is implemented by mapping the identification of the enterprise's requirements and then addressing an electronic survey to manage the requirements. The second phase involves the establishment of evaluation criteria and weights based on the results of the QFD. Weights analysis leads to decision-making, and a multi-attribute decision-making (MADM) process is implemented to ascertain the most optimal concept. Ultimately, the MADM process yields a ranking of the alternatives and determines the optimal option for the subsequent phase of the enhancement program that contributes to the continued existence and effectiveness of small manufacturing systems institutions in terms of their ability to adapt to market conditions.

Keywords: data analytics; information technology; house of quality; software; manufacturing

1 Introduction

The ongoing existence and continuation of small manufacturing systems enterprises depend on their capacity to adapt to market conditions, including supply, demand, technical advancements, and competition; this is especially true for companies involved in mechanical subcontracting. In this scenario, the company's primary reliance is on the consumer, and its success is contingent upon the customer's selection of products or services [1]. The organization must design strategies and protocols to regulate consumer inquiries, manufacturing operations, and the pursuit of customer satisfaction [2]. Additionally, it should focus on information technology, data management, and analysis, the most immediate avenues for progress [3]. Subsequently, a computer application is developed that utilizes a model employed by HoQ for requirement analysis and data analysis. This application generates results and recommendations about the product or service, facilitating informed decision-making and enhancing quality improvement. This is accomplished through the process of extrapolating customer data by identifying customer preferences and choices and then conducting experiments using this data on the proposed model within a manufacturing organization as anticipated data and decisions. HoQ, chosen as a paradigm for decision-making, utilized modified software to fulfill the specific requirements of HoQ and analyze the data derived from the resolution. In this research, the aim was to reach an enhanced decision by integrating HoQ with the data analytics. It is important to acknowledge that the application may not be suitable for big manufacturing firms due to the intricate nature of their components and operations, as well as the challenge of incorporating human behavior into automated processes within manufacturing organizations [4].

2 Related Work

2.1. QFD

QFD is a helpful decision-making technique and systematic tool that translates customer desires into specific design requirements utilized in diverse sectors to enhance customer satisfaction and product quality [5]. Research has shown that QFD can enhance e-banking operations by pinpointing crucial quality elements and conducting sensitivity analysis [6]. In maritime shipping, QFD reevaluates and rearranges the criteria for port selection; this reflects the incorporation of shipping lines into worldwide supply chains and the evolving dynamics of factors influencing port selection [7]. In lean manufacturing, using fuzzy QFD and failure mode and effects analysis (FMEA) helps to prioritize important resources and identify risks related to implementing lean tools; this leads to better resource allocation and waste reduction in manufacturing processes [9]. These studies demonstrate the wide range of applications and the usefulness of QFD in assisting decision-making processes in various operational environments. Historically, Yoji Akao developed QFD in Japan in 1966, and companies such as Mitsubishi Heavy Industries and Toyota Motor Company Ltd. later adopted it [8]. The method spread to the United States in 1983, with companies such as Xerox and Ford among the first to implement it and showcase its global impact and adoption across industries [10, 11]. QFD enables businesses, including those in the tourism industry, to align their capabilities with customer demands, improving service quality and increasing customer satisfaction. Companies can enhance their products or services by prioritizing customer requirements and implementing technical responses based on QFD analyses to better meet customer expectations and drive business success [12].

2.2. HoQ

HoQ is a systematic approach used to convert customer requirements (customer requirements) into suitable technical specifications for every phase of product development and manufacturing [13]. It is a part of the quality function deployment (QFD) process and drives the development process of a product or service. HoQ has six phases: customer requirements (whats), technical requirements (whats), relationship matrix (ROOF), correlation matrix, weights, and benchmarks [14]. Manipulating data according to these phases tends to identify real customer requirements and assist in making informed good decisions about the case, mapping at the same time the competitors' benchmarking comparisons [15].

2.3. Data Analytics

Data analytics involves examining large datasets to uncover hidden patterns, correlations, and insights. In manufacturing systems, data analytics are crucial for understanding customer requirements, improving decision-making processes, and optimizing production to meet those requirements effectively [16]. Therefore, analyzing customer feedback and purchasing patterns helps manufacturers to understand the most desired products or features. Data analytics can predict gaps and failures before they occur, minimizing downtime and saving costs even in supply chain control and optimization. Data analytics summarize historical data to understand what has happened, investigate the reasons behind past outcomes, use statistical models and machine learning techniques to predict future outcomes, and suggest actions to achieve desired outcomes, often using optimization and simulation techniques [12, 13]. More accurately, The paper predict demands to improve data analytics. Real-time monitoring helps better resource allocation, and tailoring products and services to individual customer preferences increases satisfaction and loyalty. Therefore, linking data analytics to QFD-based decision-making translates customer requirements into technical characteristics for a product, which enhances QFD data collection and analysis of customer feedback to ensure that the product features align with customer desires; this leads to understanding competitors' strengths and weaknesses to position the product effectively [17].

2.4. Manufacturing Systems

Two great approaches can be distinguished when redesigning manufacturing systems. First, the reengineering of manufacturing systems concerns the business process reengineering (BPR) philosophy. Second, the manufacturing systems' continuous improvement (CI) is related to the total quality management (TQM) philosophy. Therefore, the manufacturing systems field presents various decision-support models to facilitate the implementation of progressively higher levels of improvement. Researchers use some of the technology models, such as the algorithm for inventive problem-solving (ARIZ) combined with supervised machine learning [1], model-based systems engineering (MBSE) for designing flexible systems [2], semantic models for knowledge bases and analytical parts [3], and modeling and simulation for suggestions on how to improve performance [4].

Additionally, researchers have investigated constructing a decision support system with numerous attributes to forecast and quantify the risk of failures in complex manufacturing settings, identify workstations prone to failure, and support proactive failure avoidance [5]. By incorporating these methodologies, manufacturing systems can reap the benefits of data-driven decision-making processes, allowing them to achieve continuous improvement and operational excellence without sacrificing quality. As a result, quality function deployment (QFD) is an appropriate platform that guarantees the communication and transfer of information between the strategic, operational, and technical levels [18].

3 Utilizing the HoQ

In today’s competitive landscape, businesses face relentless pressure to align their products and services with customer expectations. Analytical tools are vital in bridging the gap between customer requirements and product features. Among these tools, HoQ stands out as a key component of QFD [19]. HoQ provides a structured methodology for translating customer requirements into precise engineering specifications, ensuring organizations can effectively respond to market demands. This paper explores how analytical analysis using the HoQ framework aids organizations in making informed decisions that optimize product development and improve customer satisfaction [14]. HoQ operates as a matrix-based tool encompassing several critical components, each contributing to a comprehensive understanding of customer requirements. Firstly, it captures customer requirements (CRs), which detail “what” customers desire. Secondly, it defines technical requirements (TRs) that outline “how” an organization plans to fulfill these requirements [20]. The relationship matrix maps the strength of the connection between CRs and TRs, allowing teams to visualize the interplay between customer expectations and technical capabilities.

Additionally, HoQ incorporates competitive benchmarking, which compares product performance against that of competitors, fostering an understanding of market positioning. Finally, the prioritization component identifies critical areas for improvement based on both customer importance and technical feasibility, guiding focused efforts to enhance product offerings [21]. Analytical analysis significantly enhances the effectiveness of the HoQ framework [22]. It begins with a customer-centric focus, emphasizing the need to gain insights into customer requirements through surveys, interviews, and thorough market analysis. Techniques like statistical sampling or clustering can identify patterns, thereby prioritizing the most critical requirements [23].

Furthermore, analysts can use the relationship matrix to evaluate the correlations between customer requirements and technical features. Analytical tools, including correlation coefficients and regression analysis, identify conflicts or synergies essential for guiding design teams in making optimal trade-offs. Moreover, the competitive analysis facilitated by HoQ provides valuable insights into the strengths and weaknesses of competitors. Analytical techniques such as SWOT analysis or market positioning maps help to visualize potential opportunities for differentiation in the marketplace [24]. HoQ also assists in optimizing resource allocation, as weighted scores derived from the framework allow decision-makers to distribute resources effectively. Techniques such as linear programming or decision matrices can refine priorities, ensuring a balance of cost, time, and quality in product development [25].

Furthermore, HoQ facilitates scenario analysis by modeling the impacts of changes in design parameters. Organizations can predict outcomes using sensitivity analysis or simulation tools and proactively adjust their strategies to align with evolving customer expectations. By incorporating these analytical methodologies, the HoQ framework empowers businesses to streamline their product development processes, ultimately enhancing customer satisfaction and fostering long-term success in the competitive marketplace [26].

4 Analytical Analysis with HoQ Framework

The HoQ framework emphasizes a customer-centric approach, which is vital for aligning product development with what customers truly need [23]. Organizations can better understand customer desires by using various methods like surveys, interviews, and market analyses. Analytical techniques, such as statistical sampling and clustering, play a key role in identifying patterns and prioritizing these requirements, ensuring that the most important customer requirements are effectively addressed during the product development process [22–23]. Moreover, HoQ helps to identify correlations and conflicts between customer requirements and technical features. The relationship matrix within HoQ allows analysts to assess the strength of these correlations, while tools like correlation coefficients and regression analysis help to uncover potential conflicts or synergies [23]. This insight is crucial for design teams, guiding them in making informed design decisions and optimizing trade-offs. Competitive analysis is another essential aspect of the HoQ framework. By utilizing benchmarking data, organizations can gain valuable insights into their competitors' strengths and weaknesses [22]. Analytical methods like SWOT analysis or market positioning maps effectively illustrate opportunities for market differentiation [23]. This analysis helps organizations pinpoint strategic advantages that can enhance their product offerings and market presence. Additionally, the HoQ framework aids in optimizing resource allocation [27]. Weighted scores from various HoQ components allow decision-makers to allocate resources more efficiently [24]. Techniques such as linear programming or decision matrices can help refine priorities, ensuring a balanced approach to cost, time, and quality in product development efforts. This strategic allocation of resources is crucial for achieving operational efficiency. Furthermore, the HoQ framework aids in scenario analysis by simulating the effects of changes in design parameters. By employing tools like sensitivity analysis or simulation, organizations can predict outcomes across various scenarios, allowing them to make proactive adjustments to meet changing customer expectations. This ability is especially crucial in fast-paced business environments where customer requirements and technological advancements can change quickly [28].

5 Benefits of Analytical Analysis Using the HoQ

The advantages of using analytical analysis through the HoQ framework are substantial. A key benefit is the alignment with customer expectations, as systematically connecting customer requirements with technical specifications ensures this alignment throughout the development process [29]. Additionally, employing analytical methods enhances data-driven decision-making, which improves the accuracy and reliability of decisions based on quantitative insights. Identifying critical features early on not only minimizes the chances of redesigns but also speeds up the product's time-to-market, helping organizations stay competitive [30]. Furthermore, by understanding market trends and competitor performance, organizations can drive innovation and differentiation, thereby strengthening their competitive edge [22]. By utilizing the insights gained from analytical analysis, companies can strategically position themselves in the marketplace [31].

6 Challenges and Limitations

While the HoQ framework offers many advantages, there are also challenges to consider when implementing it [30]. One major issue is the complexity that arises in large-scale applications; using HoQ for intricate products often demands advanced tools and specialized knowledge. Moreover, the process of assigning weightings and evaluating relationships can be subjective, which may introduce bias into the analysis unless strong analytical validation methods are in place [32]. It is also crucial to recognize that the ever-changing market conditions can limit the effectiveness of the static matrices typically used in HoQ. Rapid shifts in customer preferences and technological progress might not be adequately reflected in these static matrices, highlighting the need for continuous updates and adjustments to the analysis. Tackling these challenges is vital for ensuring the HoQ framework is effective in practical applications [33].

7 Methods

This paper took a mixed-methods approach, combining both quantitative and qualitative data analysis to create a software application called QFDSys, aimed at enhancing decision-making processes in small and medium-sized enterprises (SMEs). The methodology was grounded in the quality function deployment (QFD) framework and the house of quality (HoQ) model.

For data collection and requirements, the paper collecting data through an electronic survey sent out to the maintenance staff of a prominent manufacturing company. This survey, which included 82 questions, zeroed in on the key enablers and outcomes of an effective maintenance model. Over the course of two months, fifty-six valid responses were gathered. This quantitative data laid the groundwork for understanding customer requirements and preferences regarding maintenance processes.

To Implement the HoQ Model, the collected survey data was then analyzed using the HoQ framework implemented within the QFDSys application. This involved several key steps:

- Mapping customer requirements (whats): The first step involved determining and ranking the requirements and wants of the customers, or what they required from the maintenance services. To figure out the most important elements, the survey responses were categorized and weighted.

- Defining technical requirements (hows): The next step translated these customer requirements into technical requirements (TRs) by the means of how the company could fulfill these requirements through its operational processes.

- Creating relationships matrix: The relationship between CRs and TRs was established, quantifying the strength of the connection between customer desires and technical capabilities.

- Formulating competitive benchmarks: To compare the company's products with those of its rivals, competitive benchmarking data was included to the HoQ matrix. This made it possible to identify prospective benefits and drawbacks.
- Allocation of resources: Critical areas for improvement were determined using the HoQ's weighted scores, and resources were effectively distributed to meet the most pressing requirements.

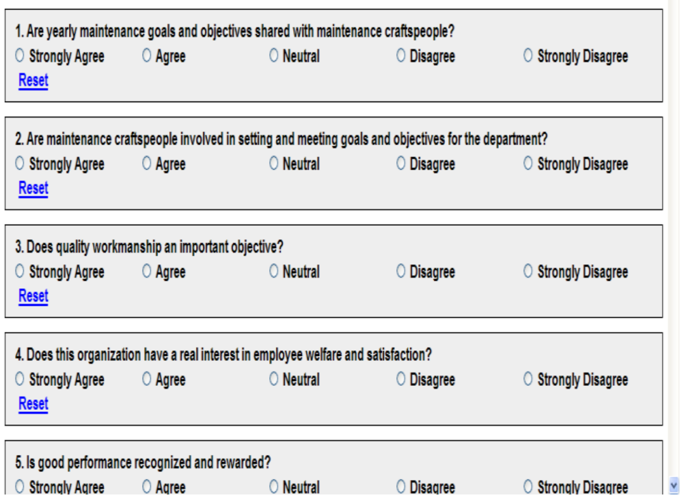
The data from the HoQ matrix was analyzed using the QFDSys software application, which was created in a Microsoft Studio environment utilizing sophisticated programming languages. A cascaded approach of converting high-level customer wants into specific technical requirements was made possible by this program, which made it easier to create many QFD matrices. Tools for data administration, analysis, and tracking the impact of enhancements on system performance were also included in the program.

To validate and determine how well the model aligned customer requirements with technical standards, the QFDSys findings were examined and analyzed. The paper also addresses difficulties in applying the HoQ methodology to larger businesses or taking into account changing market conditions, as well as the methodology's drawbacks, such as subjective weighting and the possibility of bias.

Future research will focus on enhancing the QFDSys application, integrating it with other methodologies (Lean, Six Sigma, Agile), incorporating advancements in data analytics (AI, IoT), and broadening stakeholder engagement to improve decision-making within SMEs.

8 E-survey

An electronic survey was designed for the maintenance staff of a leading manufacturing system enterprise. This survey included 82 questions that reflected the enablers and diverse outcomes of the model of excellence in maintenance. Within two months, the results of the online questionnaire were obtained. We collected fifty-six viable responses to meet the study's objectives. We incorporated the data into the model using implementation and its matrix, subjecting it to relationships and competition to determine the most reliable action or maintenance. Figure 1 represents a view of the electronic survey.



1. Are yearly maintenance goals and objectives shared with maintenance craftspeople?
○ Strongly Agree ○ Agree ○ Neutral ○ Disagree ○ Strongly Disagree
Reset

2. Are maintenance craftspeople involved in setting and meeting goals and objectives for the department?
○ Strongly Agree ○ Agree ○ Neutral ○ Disagree ○ Strongly Disagree
Reset

3. Does quality workmanship an important objective?
○ Strongly Agree ○ Agree ○ Neutral ○ Disagree ○ Strongly Disagree
Reset

4. Does this organization have a real interest in employee welfare and satisfaction?
○ Strongly Agree ○ Agree ○ Neutral ○ Disagree ○ Strongly Disagree
Reset

5. Is good performance recognized and rewarded?
○ Strongly Agree ○ Agree ○ Neutral ○ Disagree ○ Strongly Disagree

Figure 1. A view of the electronic survey Source: QFDSys

9 QFDSys Results

QFDSys application maps the HoQ model. Therefore, this application has been allocated to manufacturing systems data to serve manufacturing environmental data and variables. HoQ includes essential inputs to customer requirements (whats), a description of the interrelationship between technical descriptions (hows), as well as the relationship between requirements and descriptions (relations), the interrelationship between technical descriptions (ROOF) representing the roof of the house, identification of important priority customer requirements, competitive analysis or market potential (important), and priority technical descriptions. Figure 2 shows the essential components of HoQ represented by QFDSys. In Figure 3, the graphical interface menu shows ten QFD matrices created for the project. The active QFD-matrix is identified by its red color (QFD2). In this case, the user can manipulate this matrix's different characteristics, passing from one phase to another. A database menu is used to create a standalone database for each phase, serving the data and characteristics of "whats," "hows," and "importance."

Furthermore, Figure 4 defines customer satisfaction, product quality, delivery techniques, and manufacturing costs. Figure 5 presents improved production processes, material types, manufacturing maintenance techniques, and effective supply chain management. Figure 6 presents the relationship between customer requirements and technical requirements. Figure 7 identifies the relationships between technical requirements and their impact on each other, and Figure 8 presents a legend (symbol) of the important degrees of the relationships between whats and hows.

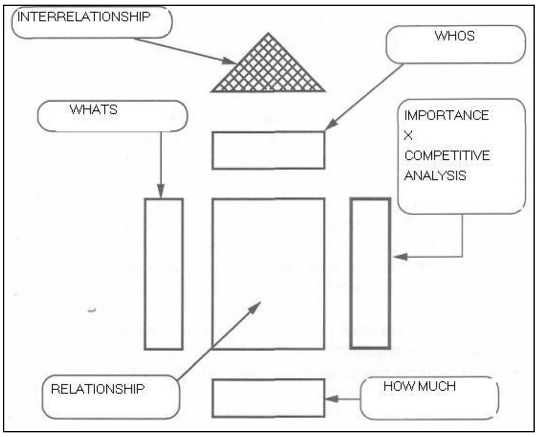


Figure 2. Main components of the house of quality (HoQ)

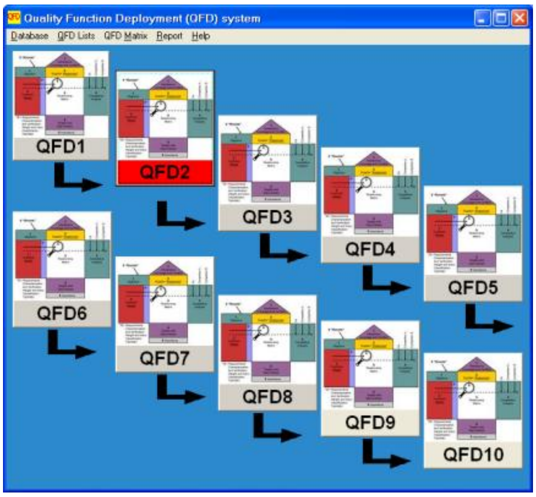
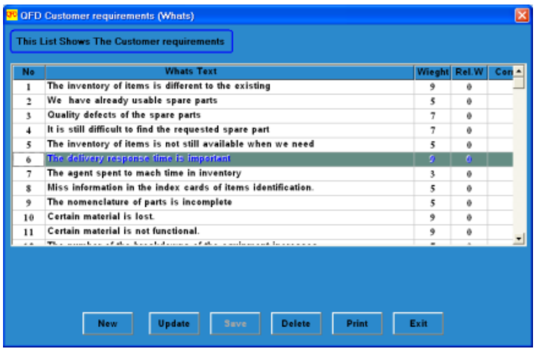


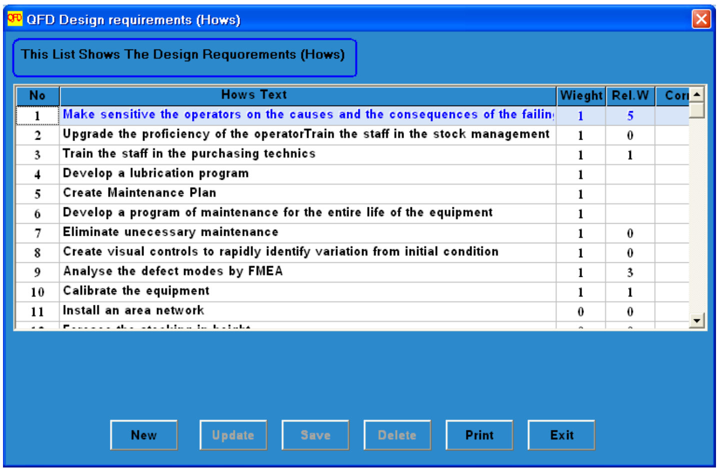
Figure 3. Cascade of QFD charts, involving a series of linked matrices that translate high-level customer requirements into detailed technical requirements



This List Shows The Customer requirements

No	Whats Text	Wieght	Rel.W	Con
1	The inventory of items is different to the existing	9	0	
2	We have already usable spare parts	5	0	
3	Quality defects of the spare parts	7	0	
4	It is still difficult to find the requested spare part	7	0	
5	The inventory of items is not still available when we need	5	0	
6	The delivery response time is important	9	0	
7	The agent spent to mach time in inventory	3	0	
8	Miss information in the index cards of items identification.	5	0	
9	The nomenclature of parts is incomplete	5	0	
10	Certain material is lost.	9	0	
11	Certain material is not functional.	9	0	

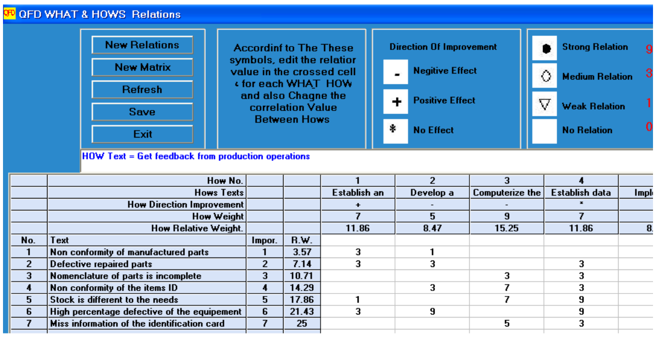
Figure 4. Customer requirements (whats)



This List Shows The Design Requorements (Hows)

No	Hows Text	Wieght	Rel.W	Cori
1	Make sensitive the operators on the causes and the consequences of the failin	1	5	
2	Upgrade the proficiency of the operatorTrain the staff in the stock management	1	0	
3	Train the staff in the purchasing technics	1	1	
4	Develop a lubrication program	1		
5	Create Maintenance Plan	1		
6	Develop a program of maintenance for the entire life of the equipment	1		
7	Eliminate unecessary maintenance	1	0	
8	Create visual controls to rapidly identify variation from initial condition	1	0	
9	Analyse the defect modes by FMEA	1	3	
10	Calibrate the equipment	1	1	
11	Install an area network	0	0	

Figure 5. Technical requirements (hows)



Accordinf to The These symbols, edit the relatior value in the crossed cell ‹ for each WHAT HOW and also Chagne the correlation Value Between Hows

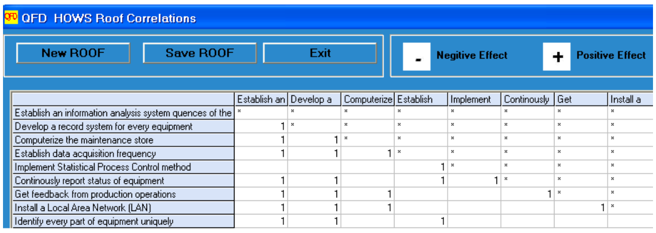
-	Negitive Effect
+	Positive Effect
*	No Effect

●	Strong Relation	9
○	Medium Relation	3
▽	Weak Relation	1
	No Relation	0

HOW Text = Get feedback from production operations

	How No.			1	2	3	4	
	Hows Texts			Establish an	Develop a	Computerize the	Establish data	Impl
	How Direction Improvement			+	-	-	*	
	How Weight			7	5	9	7	
	How Relative Weight.			11.86	8.47	15.25	11.86	8.
No.	Text	Impor.	R.W.					
1	Non conformity of manufactured parts	1	3.57	3	1			
2	Defective repaired parts	2	7.14	3	3		3	
3	Nomenclature of parts is incomplete	3	10.71			3	3	
4	Non conformity of the items ID	4	14.29		3	7	3	
5	Stock is different to the needs	5	17.86	1		7	9	
6	High percentage defective of the equipement	6	21.43	3	9		9	
7	Miss information of the identification card	7	25			5	3	

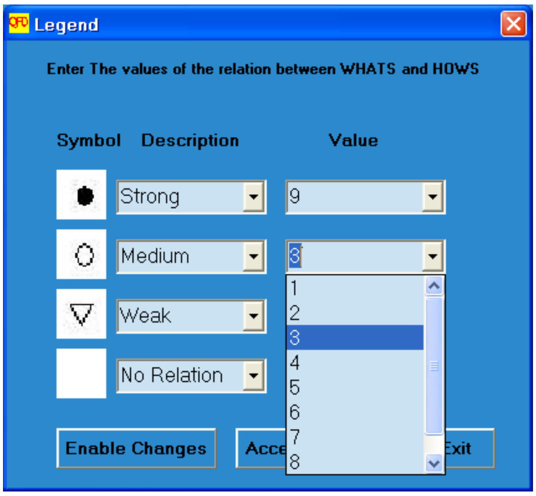
Figure 6. Matrix of relationships



-	Negitive Effect
+	Positive Effect

	Establish an	Develop a	Computerize	Establish	Implement	Continously	Get	Install a
Establish an information analysis system quences of the	*	*	*	*	*	*	*	*
Develop a record system for every equipment	1	*	*	*	*	*	*	*
Computerize the maintenance store	1	1	*	*	*	*	*	*
Establish data acquisition frequency	1	1	1	*	*	*	*	*
Implement Statistical Process Control method				1	*	*	*	*
Continously report status of equipment	1	1		1	1	*	*	*
Get feedback from production operations	1	1	1			1	*	*
Install a Local Area Network (LAN)	1	1	1				1	*
Identify every part of equipment uniquely	1	1		1				

Figure 7. Correlation matrix



Enter The values of the relation between WHATS and HOWS

Symbol	Description	Value
●	Strong	9
○	Medium	3
▽	Weak	
	No Relation	

1
2
3
4
5
6
7
8

Figure 8. Assessment of importance

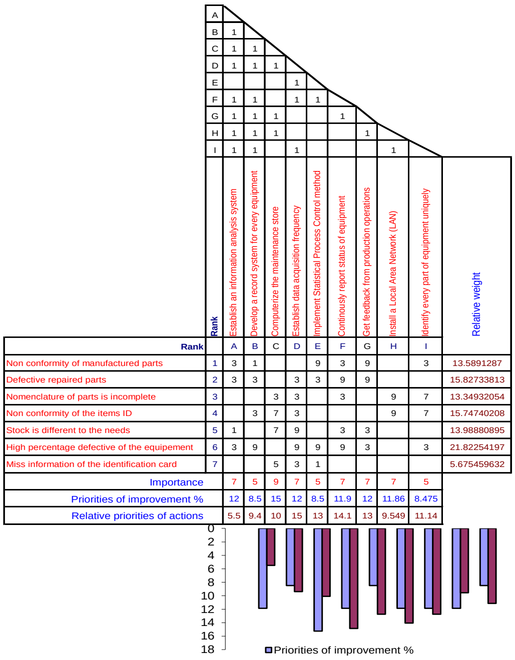
By implementing QFDSys, we hope that HoQ matrix can achieve significant enhancements in the studied case. Furthermore, the effects of improvement measures on the system's performance and quality can be consistently monitored and evaluated, leading to the establishment of a high-quality manufacturing system that meets the requirements of both external and internal users. Additionally, this approach aims to enhance production efficiency and boost the competitiveness of small and medium-sized enterprises in particular.

10 Gaps in the Proposed Methodology

The paper takes a closer look at the challenges a proposed model faces that aims to weave human behavior into the automated processes of large manufacturing companies. Some of the main hurdles include the complicated methodology, the subjective nature of how weights are assigned, and the rigid structure of the matrices used. Additionally, the limited sample size of the survey might not truly reflect the broader market trends. The questions chosen for the survey and the level of engagement may not fully address what customers need, and technological constraints could hinder its effectiveness. The survey might miss insights from other important stakeholders by focusing primarily on maintenance staff. Plus, relying on historical data for competitor comparisons may not accurately capture the current competitive landscape. Lastly, the project's plan for ongoing improvement might overlook other vital areas of business enhancement such as boosting supply chain efficiency and investing in employee training.

11 Conclusion

The HoQ technique was applied to the e-survey responses, allowing us to track the matrices of HoQ by first identifying and evaluating customer requirements based on their expected requirements. We determined the technical specifications from the statements and observations of maintenance technicians. We gathered competitor comparisons from previous studies by reviewing specifications and estimating requirements in a competitor matrix. Using the application QFDSys, which served as a repository for all data according to the matrices required by the HoQ model, we estimated the relationships between customer requirements and technical specifications, including the strength of these relationships and the correlations among the specifications themselves. As a result, we obtained values representing the weights of these relationships and links. We also identified the general trend through the competitor matrix, which enabled us to make informed decisions based on computer analysis. This improved decision-making process can be relied upon when estimating the type and method of maintenance required. Figure 9 illustrates these results. The research objectives were met by achieving effective and improved maintenance decisions while also considering how to outperform competitors. Additionally, we verified the requirements' validity and relationships, ensuring they closely align with customer expectations. In detailing these results, we observed that some specifications, such as productivity and direct technical support, do not fully meet the requirements, highlighting gaps in productivity and acceptable safety procedures and technical support.



	A	B	C	D	E	F	G	H	I
A									
B	1								
C	1	1							
D	1	1	1						
E				1					
F	1	1		1	1				
G	1	1	1			1			
H	1	1	1				1		
I	1	1		1				1	

	Rank	Establish an information analysis system	Develop a record system for every equipment	Computerize the maintenance store	Establish data acquisition frequency	Implement Statistical Process Control method	Continously report status of equipment	Get feedback from production operations	Install a Local Area Network (LAN)	Identify every part of equipment uniquely	Relative weight
Rank		A	B	C	D	E	F	G	H	I	
Non conformity of manufactured parts	1	3	1			9	3	9		3	13.5891287
Defective repaired parts	2	3	3		3	3	9	9			15.82733813
Nomenclature of parts is incomplete	3			3	3		3		9	7	13.34932054
Non conformity of the items ID	4		3	7	3				9	7	15.74740208
Stock is different to the needs	5	1		7	9		3	3			13.98880895
High percentage defective of the equipement	6	3	9		9	9	9	3		3	21.82254197
Miss information of the identification card	7			5	3	1					5.675459632
Importance		7	5	9	7	5	7	7	7	5	
Priorities of improvement %		12	8.5	15	12	8.5	11.9	12	11.86	8.475	
Relative priorities of actions		5.5	9.4	10	15	13	14.1	13	9.549	11.14	

Figure 9. Technical improvement matrix-based maintenance

12 Future Work

The QFDSys model is used as a quantitative decision-making tool for further development. The model aims to enhance manufacturing quality and decision-making by leveraging different manufacturing institutions from different industries. This model will also explore dynamic house of quality (HoQ) matrices that can adapt to changing market conditions and evolving customer requirements. In addition, it will gather insights from a broader range of stakeholders, including customers, suppliers, and management. Developing this system aligns with the research objective to seek to improve weight assignment methods and add new techniques, such as those used in the analytical hierarchy process, to minimize any subjectivity involved or adjust bias. Of course, the model requirements a training program to enhance SMEs' technical skills in applying the QFDSys approach. The QFD methodology will, therefore, be applied in different and broader areas, including product development, service quality improvement, and supply chain management by providing a framework for continuous improvement and integrating QFDSys with Lean, Six Sigma, or Agile methodologies. In addition, integrating the latest technologies, such as the internet of things (IoT) and artificial intelligence (AI), will be examined to improve data collection, analysis efficiency, and decision-making processes.

13 Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

14 References

[1] J. Díaz-Arancibia, J. Hochstetter-Diez, A. Bustamante-Mora, S. Sepúlveda-Cuevas, I. Albayay, and J. Arango-López, "Navigating digital transformation and technology adoption: A literature review from small and medium-sized enterprises in developing countries," Sustainability, vol. 16, no. 14, p. 5946, Jan. 2024, doi: https://doi.org/10.3390/su16145946.

[2] N. Rane, A. Achari, and S. P. Choudhary, "Enhancing customer loyalty through quality of service: Effective strategies to improve customer satisfaction, experience, relationship, and engagement," International Research Journal of Modernization in Engineering Technology and Science, vol. 5, no. 5, pp. 427–452, May 2023, doi: https://doi.org/10.56726/irjmets38104.

[3] R. Clancy, D. O'Sullivan, and K. Bruton, "Data-driven quality improvement approach to reducing waste in manufacturing," The TQM Journal, vol. ahead-of-print, no. ahead-of-print, Aug. 2021, doi: https://doi.org/10.1108/tqm-02-2021-0061.

[4] F. Koay, Choo, S.Teh, P. Teoh, and H. Low, "Supporting decision making with an ARIZ-based model for smart manufacturing," Malaysian Journal of Computer Science, vol. 36, no. 1, pp. 53–78, Jan. 2023, doi: https://doi.org/10.22452/mjcs.vol36no1.4.

[5] H. Elhegazy, A. Ebid, I. Mahdi, S. Haggag, and I. Abdul-Rashied, “Implementing QFD in decision making for selecting the optimal structural system for buildings,” Construction Innovation, vol. 21, no. 2, pp. 345–360, Sep. 2020, doi: https://doi.org/10.1108/CI-12-2019-0149

[6] K. Tliba, O. Penas, T. M. L. Diallo, R. Ben Khalifa, N. Ben Yahia and J. -Y. Choley, "Model Based Systems Engineering approach for the improvement of manufacturing system flexibility," 2020 21st International Conference on Research and Education in Mechatronics (REM), Cracow, Poland, 2020, pp. 1-6, doi: 10.1109/REM49740.2020.9313871.

[7] A. Sohr, F. G. Listl, K. Ecker, J. Fischer, J. C. Wehrstedt, and M. Weyrich, “Decision modeling for an ISA-95 based production ontology,” IFAC-PapersOnLine, vol. 55, no. 10, pp. 371–376, Jan. 2022, doi: https://doi.org/10.1016/j.ifacol.2022.09.421.

[8] N. Abdulatif, S. Yasser, I. Fahim, Y. Emad, A. Saleh, and S. Kassem, “Decision support using simulation to improve productivity: A case study,” 2021 International Conference on Decision Aid Sciences and Application (DASA), pp. 1120–1127, Nov. 2020, doi: https://doi.org/10.1109/dasa51403.2020.9317043.

[9] J. Sun, H. Wang, and Z. Cui, “Alleviating the Bauxite Maritime supply chain risks through resilient strategies: QFD-MCDM with intuitionistic fuzzy decision approach,” Sustainability, vol. 15, no. 10, pp. 8244–8244, May 2023, doi: https://doi.org/10.3390/su15108244.

[10] G. Paltayian, A. Georgiou, and K. Gotzamani, “A combined QFD-AHP decision-making tool for the investigation and improvement of e-banking usage,” International Journal of Quality & Reliability Management, vol. 41, no. 1, pp. 150–172, May 2023, doi: https://doi.org/10.1108/ijqrm-02-2021-0030.

[11] G. Kou, H. Dinçer, S. Yüksel, and F. S. Alotaibi, “Imputed expert decision recommendation system for QFD-based omnichannel strategy selection for financial services,” International Journal of Information Technology & Decision Making, Mar. 2023, doi: https://doi.org/10.1142/s0219622023300033

[12] H. N. Cahya and Z. S. Dila, “Penerapan quality function deployment Untuk Meningkatkan EServequal (Implementation of quality function deployment to improve E-Servequal),” Akutansi Bisnis & Manajemen (ABM), vol. 29, no. 1, p. 1, Apr. 2022, doi: https://doi.org/10.35606/jabm.v29i1.1020.

[13] J. L. Bossert, “Quality Function Deployment: The Practitioner's Approach,” Jul. 2021, CRC Press. doi: https://doi.org/10.1201/9781003066545.

[14] S. A. Kusumastuti, T. Kusmantini, and S. Sabihaini, “Analysis of service quality design with integration of Kano model and house of quality (HoQ),” Pattimura Proceeding: Conference of Science and Technology, pp. 56–70, 2023, Accessed: Feb. 11, 2025. [Online]. Available: https://ojs3.unpatti.ac.id/index.php/pcst/article/view/9273

[15] L. A. Cox, “Data analytics and modeling for improving decisions,” International Series in Management Science/Operations Research, pp. 37–64, Jan. 2023, doi: https://doi.org/10.1007/978-3-031-32013-2_2.

[16] A. Sayal, “Data analytics,” Advances in systems analysis, software engineering, and high performance computing book series, pp. 1–27, Sep. 2022, doi: https://doi.org/10.4018/978-1-6684-5722-1.ch001.

[17] T. H. A. Bahia, A. R. Idan, and K. R. Athab, “The effect of quality function deployment (QFD) in enhancing customer satisfaction,” International Journal of Professional Business Review, vol. 8, no. 1, pp. e01156–e01156, Jan. 2023, doi: https://doi.org/10.26668/businessreview/2023.v8i1.1156.

[18] R. Wolniak and W. Grebski, “The usage of quality function deployment in industry 4.0 conditions,” Scientific Papers of Silesian University of Technology. Organization and

Management Series, vol. 2023, no. 185, pp. 583–594, 2023, doi: https://doi.org/10.29119/1641-3466.2023.185.33.

[19] D. S. Taptajani, M. S. Mubarok, R. A. Aziz, and U. S. Hamidah, “House of quality (Hoq) approach to women’s backpack design using quality function deployment (Qfd) method,” Journal of Humanities and Social Studies, vol. 8, no. 1, pp. 264–267, 2024, Accessed: Feb. 10, 2025. [Online]. Available: https://journal.unpak.ac.id/index.php/jhss/article/view/9449/4956

[20] S. Lu, J. Zhou, and J. Ren, “Alleviating energy poverty through renewable energy technology: An investigation using a best-worst method-based quality function deployment approach with interval-valued intuitionistic fuzzy numbers,” International Journal of Energy Research, vol. 2023, p. e8358799, Feb. 2023, doi: https://doi.org/10.1155/2023/8358799.

[21] C.-F. Wu, C.-J. Wu, and H.-J. Chen, “Integrated process for developing and selecting more or less better products based on QFD and MCGP,” Mathematical Problems in Engineering, vol. 2023, no. 1, Jan. 2023, doi: https://doi.org/10.1155/2023/7543739.

[22] A. Adesina, V. Iyelolu, and P. Okpeke, “Leveraging predictive analytics for strategic decision-making: Enhancing business performance through data-driven insights,” World Journal of Advanced Research and Reviews, vol. 22, no. 3, pp. 1927–1934, Jun. 2024, doi: https://doi.org/10.30574/wjarr.2024.22.3.1961.

[23] I. Bajaj, “Customer loyalty and brand satisfaction: A study of customer centric practices,” International Journal for Multidisciplinary Research, vol. 5, no. 5, Oct. 2023, doi: https://doi.org/10.36948/ijfmr.2023.v05i05.7129.

[24] R. W. Puyt, F. B. Lie, and C. P. M. Wilderom, “The origins of SWOT analysis,” Long Range Planning, vol. 56, no. 3, p. 102304, Jun. 2023, doi: https://doi.org/10.1016/j.lrp.2023.10230.

[25] S. Ihde, L. Pufahl, M. Völker, A. Goel, and M. Weske, “A framework for modeling and executing task-specific resource allocations in business processes,” Computing, Jun. 2022, doi: https://doi.org/10.1007/s00607-022-01093-2.

[26] House, “Create a House of Quality (HoQ) aka Product Planning Matrix (PPM) for a Quality Function Deployment (QFD),” TeX - LaTeX Stack Exchange, Mar. 05, 2019. https://tex.stackexchange.com/questions/477795/create-a-house-of-quality-hoq-aka-product-planning-matrix-ppm-for-a-quality (accessed Nov. 04, 2024).

[27] C. Challoumis, “The modern vector of the development of science,” 2024. Available: https://conference-w.com/wp-content/uploads/2024/10/USA.P-0304102024.pdf#page=191

[28] K. Burghouts and J.-C. Engineering, “The technical and commercial impact of engineering changes in the conceptual housebuilding industry: A case study at MorgenWonen.” Accessed: Feb. 10, 2025. [Online]. Available: http://essay.utwente.nl/97192/1/Burghouts_MA_ET_3.pdf

[29] B. İ. Selamoğlu and Y. Kuvvetli, “A new holistic risk analysis approach based on the house of quality,” Çukurova Üniversitesi Mühendislik Fakültesi Dergisi, pp. 265–280, Mar. 2023, doi: https://doi.org/10.21605/cukurovaumfd.1273825.

[30] E. M. Achieng, “Evaluating the effect of implementing the house of quality framework on performance of small and medium enterprises (SMES) in Kenya,” African Journal of Emerging Issues, vol. 6, no. 22, pp. 70–83, 2024, Accessed: Feb. 11, 2025. [Online]. Available: https://www.ajoeijournals.org/sys/index.php/ajoei/article/view/738

[31] A. H. Syaifuddin, R. Ambarwati, and D. K. Sari, “Enhancing consumer loyalty and market competitiveness: Approaching IPA-

QFD in product development," BASKARA Journal of Business and Entrepreneurship, vol. 7, no. 1, p. 1, Oct. 2024, doi: https://doi.org/10.54268/baskara.v7i1.22534

[32] A. M. Idrees, A. I. ElSeddawy, and M. Ossama, "Knowledge discovery based framework for enhancing the house of quality," International Journal of Advanced Computer Science and Applications, vol. 10, no. 7, 2019, doi: https://doi.org/10.14569/ijacsa.2019.0100745.

Galal Eldin Abbas Eltayeb is an Assistant Professor of information technology at the Department of Management Information Systems (MIS), College of Business and Economics (CBE), Qassim University (KSA). He graduated with a bachelor's degree in computer science and statistics from Zagazig University, a master's degree in computer science from Khartoum University, and a Ph.D. in information technology from Al-Neelain University. His research interests include data analysis, AI data applications, and e-learning. Since 1993, he has held numerous administrative and academic positions in higher education institutions (E-mail: g.eltayeb@qu.edu.sa; ORCID: 0000-0003-3778-2061).

Optimizing Code Generation Efficiency Using the Polyhedral Model

SACI Abdallah*
Computer Science Department, University of BATNA 2, Algeria.
SEGHIR Rachid †
Computer Science Department, University of BATNA 2, Algeria.

Abstract

Optimizing scientific programs is critical for enhancing performance in modern computing systems, particularly in applications with stringent resource constraints such as embedded systems and parallel computing environments. In this context, the polyhedral model techniques have allowed to significantly advance the field of affine-loop-nest code generation by effectively leveraging parallelism and optimizing data locality. The present work proposes a novel approach based on the Maximal Parametric Inner-Box (MPIB) approximation algorithm, which shows promise in optimizing code generation performance. The basic idea is to introduce a new MPIB-driven transformation of the CLooG's mathematical representation of the source code, aiming at reducing the costly function calls generated during loop traversal. This leads to a significant enhancement in code performance, particularly evident with larger parameter values, where gains of up to 20% are achievable in certain cases. The preliminary results highlight notable improvements in execution time over existing techniques.

Key Words: Code Generation; Polyhedral Model; Code Optimization and Parallelization; Parametric Inner-Box; CLooG.

1 Introduction

Compiling and optimizing computer programs are of crucial importance to maximize the hardware-resource utilization and to enhance the application performances. Generating optimized code for nested loops is one of the most challenging and significant tasks in the field of code generation. This area continues to evolve, driven by the relentless pursuit of optimal performance and efficiency in modern computing systems including embedded systems that frequently operate under strict resource constraints and demand high performance for real-time applications. Additionally, the rise of parallel computing has necessitated the development of methods that can efficiently exploit parallelism inherent in loop nests.

In this context, the polyhedral model has emerged as a powerful framework, offering sophisticated tools for analyzing and optimizing affine loop nests. Through its mathematical foundations, this model provides a structured approach enabling automatic identification of parallelism, vectorization, cache locality improvement, and efficient code generation, which constitutes the main focus of the current investigation. The ability to automatically discover and leverage parallelism is particularly important in the era of multi-core processors and distributed computing environments, where parallel execution can lead to substantial performance gains.

Polyhedral code generation techniques, including CADGen (Continuous Automatic Differentiation Code Generator), CodeGen (Code Generator), CodeGen+ (Enhanced Code Generator), isl (Integer Set Library), and ClooG (Chunky Loop Generator)[2, 6, 5, 22, 21], have revolutionized the field of code generation for complex nested loops. These techniques enable considerable improvements in the performance of the generated code by effectively harnessing parallelism and optimizing data locality.

However, these methods often face limitations due to the computational overhead introduced by complex function calls in loop bounds, such as *min, max, ceil*, and *floor* functions. These operations, while essential for accurate bounds calculation, can become a bottleneck, particularly in performance-sensitive environments such as embedded systems and real-time applications.

In this article, we introduce a new approach for generating efficient code based on the concept of Maximal Parametric Inner-Box (MPIB). The main research question that our work addresses is: how can code generation be optimized to reduce computational load while maintaining accuracy? The key questions we seek to answer are: (i) What parametric factors significantly influence code generation efficiency? (ii) How can maximal inner-box approximation be applied to achieve this optimization?

Our method involves identifying and characterizing the parametric maximal inner box for each polyhedral set. This approach is used to explore and reorganize the loop transformation and optimization space, ensuring efficient utilization of computational resources. Consequently, we manipulate regular iteration domains to minimize expensive function calls during loop traversal whenever possible, which

[*] LaSTIC, Laboratory, Computer Science Department, University of BATNA 2, Algeria. Email: a.saci@univ-batna2.dz.

[†] LaSTIC, Laboratory, Computer Science Department, University of BATNA 2, Algeria. Email: r.seghir@univ-batna2.dz.

ultimately results in an improvement in the overall performance of the generated code. This improvement directly translates to better utilization of the limited computational resources.

The preliminary results of this work reveal that the proposed method offers significant advantages in terms of execution time compared to CLooG 0.18.4. This finding paves the way for further investigation of the impact of this new approach in the domain of nested loop optimization. Note that the MPIB approach can also be utilized in other real-world applications which are out of the focus of the present work, such as the reachability of hybrid dynamical systems, where it is crucial to know if the system can reach critical regions or stay within safe sets[19].

The remainder of this article is organized as follows: Section 2 delves into the foundational concepts of code generation using the polyhedral model, including its representation of programs and techniques for optimization. In Section 3, we present our approach, elucidating the mathematical foundations of the Maximal Parametric Inner Box (MPIB). The practical application of the MPIB approximation approach in generating efficient code is discussed in Section 4. Section 5 provides an illustrating example of our method. In section 6, we compare our approach with prior work. Finally, Section 7 presents a conclusion, summarizing our findings and highlighting avenues for future research.

2 BACKGROUND

In this section, we introduce the fundamental principles of code generation in the context of the polyhedral model. Our discussion delves into how programs are represented within this framework, emphasizing the role of CLooG tool (Chunky Loop Generator) in efficiently generating code from polyhedral representations.

2.1 Polyhedral Model

The polyhedral model is a powerful mathematical and geometrical framework for the analysis and optimization of programs through linear algebra and polyhedral geometry [5, 8]. It includes loop transformations, data restructuring, and various other techniques aimed at enhancing program performance [9, 13, 25, 24, 23, 4, 7]. This form of optimization techniques usually targets improving data locality and parallelism in code, which can greatly impact the overall efficiency of a program. Over the years, the polyhedral model has been proven successful in a wide range of cases and has become a fundamental tool in program optimization.

The polyhedral model proceeds through three primary steps. Initially, it expresses the original code into a geometric representation, associating each statement with a set of polyhedra. Next, it performs geometric transformations within this representation. Finally, it translates the set of polyhedra back into generated code. In this article, we focus on this later step where we target generating an efficient code based on our maximal parametric inner-box approach presented in section 3.

2.1.1 Polyhedral representation of programs

In the polyhedral model, a program is represented by an iteration domain, which is a set of affine functions mapping each statement (or point) in the original code to a point in the iteration domain [1, 16, 18].

Kuck [11] showed that the iteration domain of a loop nest (a set of nested loops), with affine lower and upper bounds, can be described by a polyhedron bounded by a set of half-spaces. Each half-space corresponds to a lower or upper bound on an index. The dimension of the polyhedron thus defined is equal to the depth of the loop nest (the number of its indices). Finally, each point with integer coordinates (integer vector) inside the polyhedron corresponds to an iteration of the loop nest. When the number of iterations is not fixed (cannot be determined at compile time), we refer to parametric loop nests. These are loop nests that contain symbolic constants (parameters) in the affine expressions of their bounds. A loop nest where all instructions are at the innermost level is called perfect. The general form of a perfect loop nest of depth d is:

```
for $i_1$=$l_1(p)$ to $u_1(p)$
  for $i_2$=$l_2(i_1,p)$ to $u_1(i_1,p)$
    ....
        for $i_d$=$l_d(i_1,i_2, \dots\ i_{d-1,p})$ to $u_d(i_1,i_2, \dots\ i_{d-1},p)$
            ....
```

where i_j (j = 1,..., d) are the indices of the loop nest, p is a parameter vector, and l_j, u_j (j = 1,..., d) are affine functions. When the loop nest is not perfect, instructions can appear at any depth level.

2.1.2 Example

Consider the following piece of code (loop nest):

```
for(i=1; i≤ n; i++)
  for(j=1; j ≤ i+m; j++)
      S(i,j);
```

The iterations of this loop nest correspond to the integer-coordinate points of the parametric polytope $P(p)$ as follows:

$$P(p) = \left\{ \binom{i}{j} \in \mathbf{Q}^2 \mid 1 \leq i \leq n \ \wedge \ 1 \leq j \leq i+m \right\}$$

where $p = [n\ m]$ is an integer parameter vector. The graphical representation of the iterations of this loop nest, when $p =$ $[n\ m]$=[5 2], is shown in Figure 1.

2.2 Code generation using the polyhedral model

Code generation has seen significant development thanks to the algorithm introduced by Quilleré et al.[15]. Since then, many research efforts have been conducted to enhance the quality of the generated code [2, 10, 14, 17, 20].

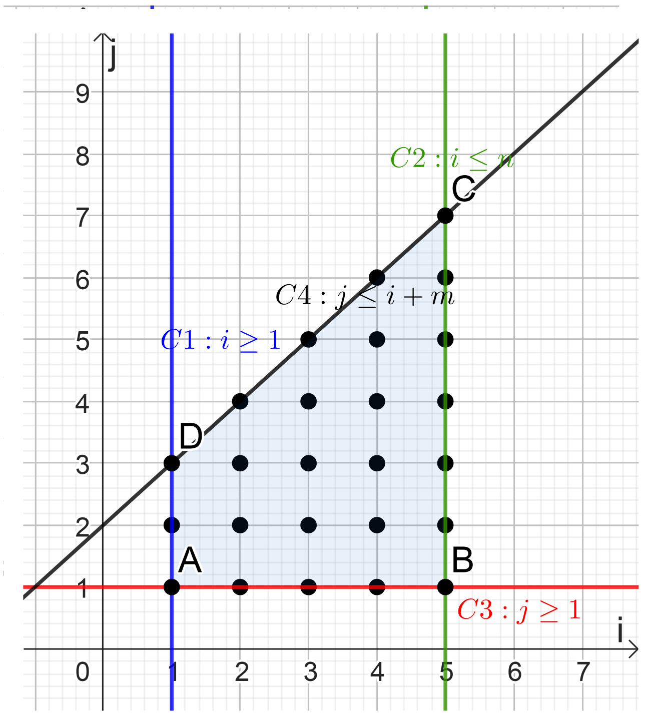


Figure 1: Representation of loop nest iterations for example 1 (with *n*=5 and *m*=2).

Quilleré's algorithm involves calculating a disjoint union of polyhedra at each recursion level across dimensions, and then generating code for each resulting subset sequentially.

Although this approach results in more extensive output code, it lowers execution complexity. This reduction is crucial for minimizing energy consumption and optimizing performance in applications with strict resource limitations, like embedded systems. However, some tests and multiple loop bounds requiring calls to *ceil/floor* and *min/max* functions are not eliminated.

This algorithm is implemented in the widely used code generation tool CLooG (Chunky Loop Generator) [2], which incorporates various enhancements aimed at preventing large code generation. These improvements include reducing the complexity of splitting, the number of scanned subsets, and the size of the generated code, all while maintaining performance [17].

2.3 CLooG: Chunky Loop Generator

CLooG (Chunky Loop Generator) is a crucial tool in the field of polyhedral code generation. It enables the efficient translation of polyhedral representations into optimized nested loop structures, which is essential for maximizing performance in resource-constrained environments or those requiring high levels of parallelism.

In the following, we give an overview of the primary algorithm used in CLooG, as initially proposed by Cedric Bastoul [2].

The CLooG tool takes as input a union of polyhedra representing the source program. Each statement of the program is thus represented by a subset of polyhedra and a set of scheduling functions. Applying these functions to the integer points of the associated polyhedra results in a new list of polyhedra that the resulting code must scan.

According to the technique proposed by LeVerge [12], the set of integer points in a polyhedron is represented as a ZPolyhedron[1].

In order to generate a loop code, the CLooG algorithm starts by computing the projection of polyhedra at dimension ($d = 1$), subsequently separating them into an ordered list of disjoint polyhedra. It then scans this list to produce code for the outermost loops (level-one loops). These disjoint polyhedra are then projected onto the second dimension ($d = 2$) to generate code for level-two loops. CLooG iterates recursively across the remaining dimensions (levels 3, 4, ...) to generate loop codes at the corresponding levels. The detailed algorithm is given in Algorithm 1.

In step 5, the algorithm computes the lower bound and the stride for each loop level ($d \in$ 1,2, ..., n) defined by its subdomains (polyhedra). Then, it merges inner polyhedra whenever possible in step 5(b.i) in order to reduce the code size. In step 5(b.ii), the function is recursively called for the next dimension ($d+1$) by intersecting the context domain with the bounds of the currently generated loop. Step 7 involves reuniting certain point polyhedra with their host polyhedra, from which they were separated in the previous step, with the aim of minimizing the overall size of the generated code [17].

Although CLooG excels at managing polyhedral sets to produce high-performance code, it has limitations, particularly with the generated *min, max, ceil,* and *floor* function calls in loop bounds, which can become computationally expensive. In the following sections (3 and 4), we will show that our MPIB-based method reduces these costs by simplifying loop bounds, thereby reducing function calls and enhancing the overall execution time of the generated code.

3 Maximal Parametric Inner-Box (MPIB) approximation approach

In this section, we propose a new approach for approximating the maximal parametric inner box based on the method of Bemporad et al. [3]. In their work, and starting from a non-parametric polytope *P*, the authors search for two collections of boxes (*I* and *E*) such that:

- The interiors of the boxes in each collection do not overlap,
- The union of all boxes in *I* is contained in *P*.
- The union of all boxes in *E* contains *P*.

Note that in the current work, we are interested in determining only one approximate maximal inner box within a

[1] A ZPolyhedron is the intersection of an integral lattice and a polyhedron.

```
Algorithm 1: CLooG's code generation algorithm [2].
  Data: A polyhedron list ($TS_1, ..., TS_n$), a context C, the
        current dimension $d$.
  Result: Code scanning the polyhedra inside the input
          list.
  Begin
    1. Intersect each polyhedron $P_i \in T_{Si}$ with the context C.
    2. Compute the projection $P_i$ onto the outermost $d$
       dimensions for each resulting polyhedron $T_{S_i}$, and
       consider the new list where $T_{S_i}$ is replaced by $P_i$.
    3. Separate the list of resulting projections $P_i$ from step 2
       into a new list of non-overlapping polyhedra.
    4. Order each list of non-overlapping polyhedra representing
       the projection $P_i$, from step 3, in the lexicographical order.
    5. For each polyhedron $P \rightarrow (T_{Sp}, \cdots, T_{Sq})$ in the list:
       (a). Compute the stride and the lower bound by looking
            for stride constraints in the $(T_{Sp}, \cdots, T_{Sq})$ list.
       (b). While there is a polyhedron in $(T_{Sp}, \cdots, T_{Sq})$:
            (i). Merge adjacent polyhedra scanning the same
                 statements in a new list.
            (ii). Recurse for the new list with the new loop
                  context $C \cap P$ and the next dimension $d+1$.
    6. Apply steps 2 to 4 of the algorithm to the inside list in
       order to eliminate dead code, for each polyhedron P inside
       the list.
    7. Reduce code size by making all possible unions of host
       polyhedra with point polyhedral.
    8. Return the code scanning the polyhedron list.
  End.
```

2-dimensional parametric polytope $P(p)$ that has one parameter $(p = [n])^2$. This box will be used in section 4 to generate an efficient code based on CLooG Algorithm. Our method can be succinctly described as follows:

Let $P(n)$ be the parametric polytope defined by:

$$P(n) = \{X \in \mathbf{R}^2 : AX \leq Bn + b\}$$

And let :

- A^+ : be the positive matrix of A.
- A_1^+ : be the first column of A^+.
- A_2^+ : be the second column of A^+.

To determine an approximation of the maximal parametric box included in $P(n)$, we start by assigning distinct values to n in order to obtain different instances of the polytope $P(n)$. For each instance of $P(n)$, we determine the maximal inner box based on the method proposed by Bemporad et al. [3], which involves the following steps:

1. Solve LP1 to find r_1, the maximum ratio along the first dimension.
   $LP1 : r_1 = max\{r : AX + A_1^+ r \leq Bn + b\}$.
2. Solve LP2 to find r_2, the maximum ratio along the second dimension.
   $LP2 : r_2 = max\{r : AX + A_2^+ r \leq Bn + b\}$.
3. Solve LP3 to determine the scaling factor λ^* which maximizes the box dimensions while ensuring it stays within $P(n)$.
   $LP3 : \lambda^* = max\{\lambda : AX + A^+ r\lambda \leq Bn + b\}$, with $r = [r_1 r_2]$.

The solution of $LP3$ is defined by: (X^*, λ^*), with $X^* = [i^* j^*]^t$.

For a given instance of the polytope $P(n)$, i.e for a given value of n, the MPIB is defined by its two extremal non-parametric points (V_1 and V_2) such that:

- $V_1(i_{V_1}, j_{V_1})$, with : $i_{V_1} = i^*$ and $j_{V_1} = j^*$.
- $V_2(i_{V_2}, j_{V_2})$, with: $i_{V_2} = i^* + r_1.\lambda^*$ and $j_{V_2} = j^* + r_2.\lambda^*$.

These points mark the endpoints of the approximate largest box included in the considered instance of polytope $P(n)$. To compute the parametric coordinates of the extremal points of the MPIB included in the parametric polytope $P(n)$, we need to compute a regression line for each of the four coordinates. These lines are given by the following equations:

$$\begin{aligned} i_{V_1}(n) &= \alpha_1.n + \beta_1 \\ j_{V_1}(n) &= \alpha_2.n + \beta_2 \\ i_{V_2}(n) &= \alpha_3.n + \beta_3 \\ j_{V_2}(n) &= \alpha_4.n + \beta_4 \end{aligned}$$

The parametric coordinates of $V_{1(n)}(i_{V_{1(n)}}, j_{V_{1(n)}})$ and $V_{2(n)}(i_{V_{2(n)}}, j_{V_{2(n)}})$ define the parametric maximal inner box of $P(n)$. Indeed, in order to determine this box, it suffices to find its two extremal points V_1 and V_2 having the lowest, respectively highest coordinates as shown in Figure 5. Our approach of approximating the MPIB is described in Algorithm 2.

Example:
Let $P(p)$ be the parametric polytope defined by the following inequations:

$$P(p) = \begin{cases} -i + j & \leq -2 \\ i + j & \leq n \\ 2i - 6j & \leq n + 4 \\ -9i + 18j & \leq n - 2 \\ n & \geq 20 \end{cases}$$

$P(p)$ can be rewritten as follows:

$$P(p) = \{X \in \mathbf{R}^2 : AX \leq Bp + b\}, \textit{ where :}$$

$$X = \begin{bmatrix} i \\ j \end{bmatrix}, A = \begin{bmatrix} -1 & 1 \\ 1 & 1 \\ 2 & -6 \\ -9 & 18 \end{bmatrix}, b = \begin{bmatrix} -2 \\ 0 \\ 4 \\ -2 \end{bmatrix}, B = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \textit{and } p = \begin{bmatrix} n \end{bmatrix},$$

[2] Our method can be extended to address problems involving higher dimensions and additional parameters.

A^+,A_1^+, and A_2^+, are given as follows:

$$A^+ = \begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 2 & 0 \\ 0 & 18 \end{bmatrix}, A_1^+ = \begin{bmatrix} 0 \\ 1 \\ 2 \\ 0 \end{bmatrix}, and\, A_2^+ = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 18 \end{bmatrix}$$

By assigning different values to the parameter n (instantiation of $P(n)$) and solving the linear programs $LP1$, $LP2$, and $LP3$ below, we obtain the coordinates of the two extremal points V_1 and V_2 of the approximate MPIB of $P(n)$, as shown in Table 1.

$$LP1: r_1 = max \left\{ r : \begin{bmatrix} -1 & 1 \\ 1 & 1 \\ 2 & -6 \\ -9 & 18 \end{bmatrix} X + \begin{bmatrix} 0 \\ 1 \\ 2 \\ 0 \end{bmatrix} r \leq \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \end{bmatrix} n + \begin{bmatrix} -2 \\ 0 \\ 4 \\ -2 \end{bmatrix} \right\}$$

$$LP2: r_2 = max \left\{ r : \begin{bmatrix} -1 & 1 \\ 1 & 1 \\ 2 & -6 \\ -9 & 18 \end{bmatrix} X + \begin{bmatrix} 1 \\ 1 \\ 0 \\ 18 \end{bmatrix} r \leq \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \end{bmatrix} n + \begin{bmatrix} -2 \\ 0 \\ 4 \\ -2 \end{bmatrix} \right\}$$

$$LP3: \lambda^* = max \left\{ \lambda : \begin{bmatrix} -1 & 1 \\ 1 & 1 \\ 2 & -6 \\ -9 & 18 \end{bmatrix} X + \begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 2 & 0 \\ 0 & 18 \end{bmatrix} r\lambda \leq \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \end{bmatrix} n + \begin{bmatrix} -2 \\ 0 \\ 4 \\ -2 \end{bmatrix} \right\}$$

, where $r = [r_1 \; r_2]$

Then, we use Microsoft Excel solver to determine the four regression lines defining the four parametric coordinates of points $V_{1(n)}(i_{V_1}(n), j_{V_1}(n))$ and $V_{2(n)}(i_{V_2}(n), j_{V_2}(n))$ from the instances of $V_1(i_{V_1}, j_{V_1})$ and $V_2(i_{V_2}, j_{V_2})$:

$$i_{V_1}(n) = 0,384258945560599.n - 0,318353165232634$$
$$j_{V_1}(n) = 0,084104990554686.n - 0,576570768496135$$
$$i_{V_2}(n) = 0,752314971664028.n + 0,270287695130927$$
$$j_{V_2}(n) = 0,247685028335843.n - 0,270287693456269$$

Finally, the parametric inner box is defined by the two extremal points $V_{1(n)}(i_{V_1}(n), j_{V_1}(n))$ and $V_{2(n)}(i_{V_2}(n), j_{V_2}(n))$ as illustrated in Figure 5 (for n=50).

In determining the regression lines for the coordinates of the maximal parametric inner box (MPIB), we experimented with various levels of decimal precision to ensure that each calculated point remains a valid integer point within the polytope, across a wide range of values for the parameter n. Specifically, we evaluated precision levels of 6, 8, 10, 12, and 15 decimal places in the regression expressions. The results showed that a precision of at least 15 decimal places was necessary to maintain the validity of all coordinates as integer points inside the polytope for n values ranging from the initial value up to 1,000,000. With fewer than 15 decimal places, certain coordinates occasionally fell outside the bounds of the polytope, which would compromise the accuracy of the MPIB approximation. Thus, we opted for 15 decimal places in the regression expressions to ensure that the MPIB coordinates reliably represent integer points within the polytope over the full range of parameter values considered in our study.

Algorithm 2: Approximation of the Maximal Parametric Inner Box.

```
Data: A parametric polytope $P(n)$.
Result: MPIB
Begin
Let :
  $A^+$ be the positive matrix of A.
  $A_1^+$ and $A_2^+$ be the first and second columns of $A^+$,
  respectively.
Let :
  $A_{i_{V_1}}[\,], A_{j_{V_1}}[\,], A_{i_{V_2}}[\,]$ and $A_{j_{V_2}}[\,]$ be the coordinate
  arrays of points $V_1$ and $V_2$ (for different values of the
  parameter).
Step 1:
$counter \leftarrow 1$;
$n \leftarrow intilal\_value$;
while $(n \leq final\_value)$ do
    Solve $LP1 : r_1 = max\{r : AX + A_1^+ r \leq Bn + b\}$.
    Solve $LP2 : r_2 = max\{r : AX + A_2^+ r \leq Bn + b\}$.
    Solve $LP3 : \lambda^* = max\{\lambda : AX + A^+ r\lambda \leq Bn + b\}$.
     with $r = [r_1 r_2]$.
    //The solution of $LP3$ is: $(X^*, \lambda^*)$,
    //where: $X^* = [i^* \; j^*]^t$.
    // The inner box is defined by the two points
    // $V_1(i_{V_1}, j_{V_1})$ and $V_2(i_{V_2}, j_{V_2})$, where
    //$i_{V_1} = i^*, j_{V_1} = j^*, i_{V_2} = i^* + r_1\lambda^*$ and
    //$j_{V_2} = j^* + r_2\lambda^*$.
    $A_{i_{V_1}}[counter] \leftarrow i^*$;
    $A_{j_{V_1}}[counter] \leftarrow j^*$;
    $A_{i_{V_2}}[counter] \leftarrow i^* + r_1\lambda^*$;
    $A_{j_{V_2}}[counter] \leftarrow j^* + r_2\lambda^*$;
    $counter \leftarrow counter + 1$;
    $n \leftarrow n + step$; //step = 100000
end
Step 2
Determination of the four regression lines corresponding
  to the values stored in arrays:
  $A_{i_{V_1}}[\,], A_{j_{V_1}}[\,], A_{i_{V_2}}[\,],$ and $A_{j_{V_2}}[\,]$ as follows:
    $i_{V_1}(n) \leftarrow \alpha_1.n + \beta_1$;
    $j_{V_1}(n) \leftarrow \alpha_2.n + \beta_2$;
    $i_{V_2}(n) \leftarrow \alpha_3.n + \beta_3$;
    $j_{V_2}(n) \leftarrow \alpha_4.n + \beta_4$;
The approximate MPIB is defined by the two parametric
  points :
    $V_{1(n)}(i_{V_1}(n), j_{V_1}(n))$ and
    $V_{2(n)}(i_{V_2}(n), j_{V_2}(n))$.
end.
```

It is worth noting that it is possible to consider coordinates

Table 1: Coordinates of V_1 and V_2 for Example 1.

n	20	100	1000	10000	100000	150000	200000	1000000
iV_1	8	39	384	3843	38426	57639	76852	384259
jV_1	2	8	84	841	8410	12616	16821	84105
iV_2	16	76	753	7524	75232	112848	150464	752316
jV_2	4	24	247	2476	24768	37152	49536	247684

with fewer decimal places. However, an additional verification step is required to ensure that all coordinates lie within the polytope. If any coordinate does not satisfy this inclusion constraint, it is necessary to adjust the point by selecting the nearest valid coordinates within the polytope.

4 ENHANCING CODE GENERATION PERFORMANCE USING THE MPIB APPROXIMATION ALGORITHM

Methodology

Our methodology is based on the Maximal Parametric Inner-Box (MPIB) approximation algorithm, which aims to optimize code generation by reducing costly function calls in loop bounds. The approach involves three main steps:

1. Start by running the first three setps of CLooG algorithm to generate a polyhedral representation of the source code to be optimized.
2. Apply the MPIB approach to convert the polyhedral representation into a new form that enhances code performance.
3. Resume the CLooG algorithm from Step 4 to generate a new code using this later polyhedral representation.

In the following, we will show how the Maximal Parametric Inner Box (MPIB) approximation approach can be applied in generating effective code. The main objective of our work is to generate an efficient code using the polyhedral model. This code will be generated by combining CLooG algorithm with our method of approximating the MPIB presented in the previous section. This approach consists in modifying CLooG algorithm (Algorithm 1) immediately after step 3. In this new algorithm, we start by calling CLooG until step 3. Then we compute the approximate MPIB for each sub-polytope P_i obtained at step 3 and replace it with the following 5 sub-polytopes $P_{i_1}, P_{i_2}, P_{i_3}, P_{i_4}$, and P_{i_5}:

- $P_{i_1} = P_i \cup \{i < i_{V_1}(n)\}$,
- $P_{i_2} = P_i \cup \{i \geq i_{V_1}(n), i \leq i_{V_2}(n), j < j_{V_1}(n)\}$,
- P_{i_3}=$\{i\geq i_{V_1}(n), i\leq i_{V_2}(n), j\geq j_{V_1}(n), j\leq j_{V_2}(n)\}$ // the MPIB,
- $P_{i_4} = P_i \cup \{i \geq i_{V_1}(n), i \leq i_{V_2}(n), j > j_{V_2}(n)\}$,
- $P_{i_5} = P_i \cup \{i \geq i_{V_2}(n)\}$.

After the step of generating this new polyhedral representation of the code to be optimized, we resume CLooG algorithm from step 4. This means that, instead of generating the code for the polyhedral set given by CLooG, we do it for the new polyhedral representation based on the MPIB approximation approach. This approach offers the advantage of efficiently handling regular polyhedral sets, requiring only a few calls to *min/max* and *floor/ceil* functions. This optimization significantly improves the execution time of the generated code.

We note that the core factor in our work is the execution time, as this is critical in evaluating the efficiency of the generated code within the polyhedral model. Algorithm 3 summarizes our MPIB-based code generation method.

It should be noted that the method for determining the MPIB can, if necessary, be applied recursively to one or all of the sub-polytopes $P_{i_1}, P_{i_2}, P_{i_4}$ or P_{i_5} when the polytope is sufficiently large. Furthermore, this method can be generalized to higher dimensions, which involves solving $(d+1)$ linear programs for a dimension d.

5 ILLUSTRATING EXAMPLE

Consider the following parametric polytope:

$$P(p) = \begin{cases} -i+j & \leq -2 \\ i+j & \leq n \\ 2i-6j & \leq n+4 \\ -9i+18j & \leq n-2 \\ n & \geq 20 \end{cases}$$

The corresponding code generated by CLooG-0.18.4 for this polytope and its graphical representation are shown in Figures 2 and 3, respectively. Note that the presence of calls to the *min/max* and *floor/ceil* functions in the inner-loop bounds of the generated code results in a substantial control overhead, affecting execution time. The idea of our approach is to avoid, as much as possible, costly function calls inside loop bounds using the modified GLooG algorithm (Algorithm 3) based on the MPIB approximation method (Algorithm 2). The resulting optimized code and its corresponding iteration domain are shown in Figures 2 and 3, respectively.

6 COMPARISON WITH PRIOR WORK AND ANALYSIS

In this section, we provide a comparative analysis between our method and existing techniques, particularly focusing on the latest version of the CLooG tool (0.18.4), which has been

```
for (i=2; i≤ floord(7*n+4,8); i++) {
  for(j=max(0,ceild(2*i-n-4,6));j≤min(min(floord(9*i+n-2,18),-i+n),i-2); j++){
    S(i,j);
  }
}
```

Figure 2: Generated code by CLooG 0.18.4

Algorithm 3: Code generation using MPIB approximation approach (**Modified CLooG's algorithm**).

```
Data: a parametric polytope P(n).
Result: Generated Code
Begin
Step 1.
Execution of Steps 1, 2 and 3 of the ClooG algorithm
  (Algorithm 1) to generate a disjoint union of polytopes
  $S = \bigcup_{i=1}^{|S|} P_i$ corresponding to the union of input
  polytopes (Polyhedral representation of the source code
  to be optimized)
Step 2
Decomposition of each polytope $P_i \in S$ into a disjoint
  union of 5 sub-polytopes $P_{i_1}, P_{i_2}, P_{i_3}, P_{i_4}$, and $P_{i_5}$, where :
  For all $l, k \in \{1,2,3,4,5\}$ and $l \neq k$ : $\begin{cases} P_i = \cup_{j=1}^{5} P_{i_j} \\ P_{i_l} \cap P_{i_k} = \emptyset \end{cases}$
with:
  - $P_{i_1} = P_i \cup \{i < i_{V_1}(n)\}$,
  - $P_{i_2} = P_i \cup \{i \geq i_{V_1}(n), i \leq i_{V_2}(n), j < j_{V_1}(n)\}$,
  - $P_{i_3} = \{i \geq i_{V_1}(n), i \leq i_{V_2}(n), j \geq j_{V_1}(n), j \leq j_{V_2}(n)\}$,
  - $P_{i_4} = P_i \cup \{i \geq i_{V_1}(n), i \leq i_{V_2}(n), j > j_{V_2}(n)\}$,
  - $P_{i_5} = P_i \cup \{i \geq i_{V_2}(n)\}$.
Step 3
Resume the CLooG algorithm from Step 4 to generate
the code corresponding to the disjoint union of polytopes
generated in the previous step (Step 2). end.
```

widely used for polyhedral code generation. While CLooG is recognized for its efficiency in generating code from polyhedral representations, our method introduces the Maximal Parametric Inner-Box (MPIB) approximation, which offers notable improvements in execution time.

One of the key differences between our approach and previous work lies in the way the loop bounds are handled. Traditional methods, including CLooG, rely heavily on the use of *min/max* and *ceil/floor* function calls, which can add significant overhead in execution. In contrast, our approach minimizes these function calls by approximating the maximal inner-box, leading to reduced computational load and enhanced performance. This difference is more significant for large parameter values.

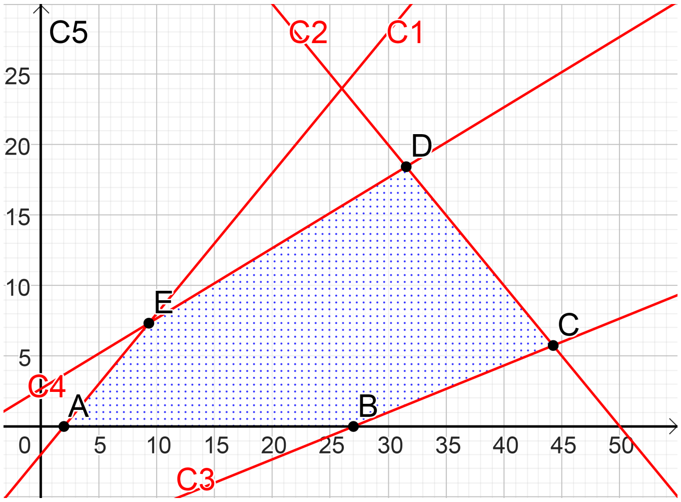


Figure 3: Graphical representation of example 1

In order to demonstrate the performance of our method, we consider the codes from Figures 2 and 4 generated by the original CLooG algorithm and our MPIB-based method respectively. These codes were compiled with gcc 5.4 and executed on an Intel i5 processor at 2.30GHz, with the values of the parameter n ranging from 100000 to 10000000 and a step of 100000.

Figure 6 and Table 2 present the execution times for both the standard CLooG algorithm and the proposed MPIB-based approach across varying parameter values. This comparison allows us to observe the substantial improvement in execution time when employing our approach, particularly noticeable with larger values of the parameter n. For instance, with n = 500000, the runtime for the code produced by CLooG-18.0.4 is 173.99 s, whereas our method yields a runtime of 140.87 s, resulting in a gain rate of 19.04%. This improvement is due to the decreased number of calls to *min/max* and *floor/ceil* functions in the code generated by our approach.

7 CONCLUSION AND FUTURE WORK

Optimizing code generation for nested loops is crucial in maximizing hardware resource utilization and enhancing application performance. The polyhedral model, with its sophisticated tools, is extensively employed in code generation algorithms.

The efficiency of the generated code is significantly impacted

```
for (i=2;i≤floord(9606473639*n-7958829131,25000000000);i++) {
   for (j=0;j≤min(floord(9*i+n-2,18),i-2);j++){
      S1(i,j);
   }
}
for(i=ceild(9606473639*n-7958828880,25000000000); i≤floord(47019685729*n+16892980945,
            62500000000);i++){
   for(j=max(0,ceild(2*i-n-4,6));j≤floord(42052495277*n-288285384248, 500000000000);j++) {
      S1(i,j);
   }
   for(j=max(ceild(2*i-n-4,6),ceild(42052495277*n-288285379248,500000000000));j≤
            floord(49537005667*n-54057538692,200000000000);j++){
      S1(i,j);
   }
   for(j=ceild(49537005667*n-54057536691,200000000000);j≤min(floord(9*i+n-2,18),-i+n);j++){
      S1(i,j);
   }
}
for(i=ceild(47019685729*n+16892981571,62500000000);i≤floord(7*n+4,8);i++) {
   for (j=ceild(2*i-n-4,6);j≤-i+n;j++) {
      S1(i,j);
   }
}
```

Figure 4: Generated code by our approach

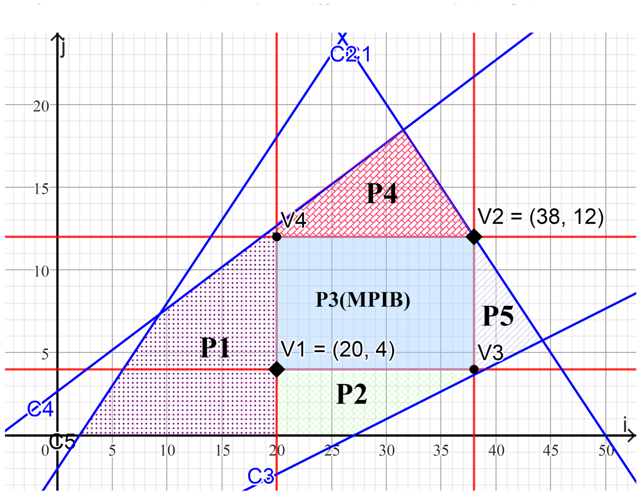


Figure 5: Separation of the polytope into 5 sub-polytopes

Table 2: Execution times for CLooG's algorithm and Our Algorithm

n	100000	200000	300000	400000	500000	600000	700000	800000	900000	1000000
CLooG(s)	6.96	27.84	62.64	111.35	173.99	250.52	340.98	445.36	564.04	689.92
MPIB(s)	5.36	21.29	50.72	90.16	140.87	202.86	276.11	360.60	456.37	563.63
Gain	1.60	6.55	11.92	21.19	33.12	47.66	64.87	84.76	107.67	126.29
Ratio(%)	22.97	23.52	19.03	19.03	19.04	19.03	19.02	19.03	19.09	18.30

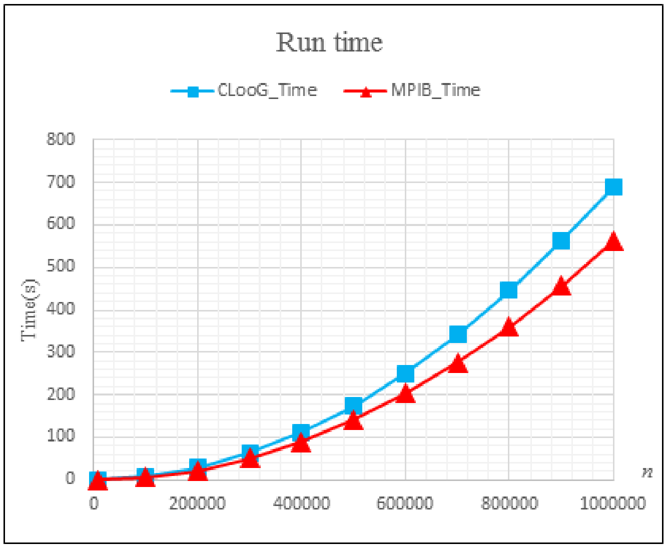


Figure 6: Execution times

by how polyhedral operations are applied to the mathematical representation of the original code.

This article presents a new approach based on the proposed parametric maximal inner-box approximation algorithm, representing a promising avenue for further enhancing code generation efficiency.

By identifying this box for each polyhedral set and performing the underling transformations, we mitigate costly function calls during loop traversal, ultimately leading a substantial improvement of the code performance, particularly with larger values of the parameter where the gain could achieve about 20% in some cases. The experimental results demonstrate notable advantages compared to existing techniques, encouraging further exploration of the potential impact of this approach in nested loop optimization.

In line with the problem statement and research contribution, our study has successfully addressed the challenge of optimizing loop-based code generation by leveraging the polyhedral model. Focusing on reducing computational overhead, particularly through the new concept of MPIB, we have provided a solution that improves execution efficiency in a way that was not previously explored.

Despite the significant improvements in execution time and efficiency achieved by the proposed MPIB-based approach, some limitations can be observed. While the method enhances performance in many cases, its effectiveness is highly dependent on the structure of the polyhedral sets being processed. In particular, when iteration domains are highly irregular or contain non-affine constraints, the approximation may not deliver optimal results. Additionally, extending the approach to higher-dimensional polyhedral sets presents challenges, as the complexity of solving additional linear programs increases significantly.

This study opens up several promising directions for future research in nested loop optimization. First, extending the proposed MPIB approach to more complex cases, such as higher-dimensional polyhedral sets dealing with deeper levels of nesting. Additionally, hybrid optimization strategies, like combining MPIB with techniques such as parametric tiling or dynamic loop transformations, could yield further performance improvements. Another key area for exploration is the integration of this method into modern compilation frameworks to facilitate its adoption by software developers and increase its usability in practical applications. Addressing these aspects would allow the proposed approach to be refined and extended to a broader range of applications in polyhedral optimization.

References

[1] A. Acharya, U. Bondhugula, and A. Cohen. “An Approach for Finding Permutations Quickly: Fusion and Dimension matching”. _eprint: abs/1803.10726. 2018.

[2] Cédric Bastoul. “Code Generation in the Polyhedral Model Is Easier Than You Think”. In: *Parallel Architectures and Compilation Techniques - Conference Proceedings, PACT*. Jan. 2004, pp. 7–16. ISBN: 0-7695-2229-7. DOI: 10.1109/PACT.2004.1342537.

[3] A. Bemporad, C. Filippi, and F. D. Torrisi. “Inner and outer approximations of polytopes using boxes”. In: *Comput. Geom.* 27 (2004), pp. 151–178.

[4] Wlodzimierz Bielecki, Marek Palkowski, and Maciej Poliwoda. “Automatic code optimization for computing the McCaskill partition functions”. In: Sept. 2022, pp. 475–478. DOI: 10.15439/2022F4. URL: https://annals-csis.org/Volume_30/drp/4.html (visited on 09/20/2024).

[5] U. Bondhugula et al. “A practical automatic polyhedral parallelizer and locality optimizer”. In: *ACM-SIGPLAN Symposium on Programming Language Design and Implementation*. 2008.

[6] Chun Chen. “Polyhedra scanning revisited”. en. In: *Proceedings of the 33rd ACM SIGPLAN Conference on Programming Language Design and Implementation*. Beijing China: ACM, June 2012, pp. 499–508. ISBN: 978-1-4503-1205-9. DOI: 10.1145/2254064.2254123. URL: https://dl.acm.org/doi/10.1145/2254064.2254123 (visited on 07/10/2024).

[7] Gianpietro Consolaro et al. “PolyTOPS: Reconfigurable and Flexible Polyhedral Scheduler”. In: *2024 IEEE/ACM International Symposium on Code Generation and Optimization (CGO)*. Edinburgh, United Kingdom: IEEE, Mar. 2024, pp. 28–40. ISBN: 9798350395099. DOI: 10.1109/CGO57630.2024.10444791. URL: https://ieeexplore.ieee.org/document/10444791/ (visited on 09/20/2024).

[8] P. Feautrier and C. Lengauer. *Polyhedron Model*. Encyclopedia of Parallel Computing, 2011.

[9] L. Gonnord et al. “A Survey on Parallelism and Determinism”. In: *ACM Computing Surveys* 55 (2022), pp. 1–28.

[10] Tobias Grosser, Sven Verdoolaege, and Albert Cohen. “Polyhedral AST Generation Is More Than Scanning Polyhedra”. en. In: *ACM Trans. Program. Lang. Syst.* 37.4 (Aug. 2015), pp. 1–50. ISSN: 0164-0925, 1558-4593. DOI: 10.1145/2743016. URL: https://dl.acm.org/doi/10.1145/2743016 (visited on 07/10/2024).

[11] D. J. Kuck. *The Structure of Computers and Computations*. Vol. 1. Wiley, 1978.

[12] H. Le Verge. “Recurrences on lattice polyhedra and their applications”. 1995.

[13] L. Narmour, T. Yuki, and S. V. Rajopadhye. “Maximal Simplification of Polyhedral Reductions”. _eprint: abs/2309.11826. 2023.

[14] L. Pouchet et al. “Iterative Optimization in the Polyhedral Model: Part I, One-Dimensional Time”. In: *International Symposium on Code Generation and Optimization (CGO’07)*. 2007, pp. 144–156.

[15] F. Quilleré, S. V. Rajopadhye, and D. Wilde. “Generation of Efficient Nested Loops from Polyhedra”. In: *International Journal of Parallel Programming* 28 (2000), pp. 469–498.

[16] W. Ranasinghe et al. “PCOT: Cache Oblivious Tiling of Polyhedral Programs”. _eprint: abs/1802.00166. 2018.

[17] H. Razanajato, V. Loechner, and C. Bastoul. “Splitting polyhedra to generate more efficient code”. In: *IMPACT 2017, 7th International Workshop on Polyhedral Compilation Techniques*. Jan. 2017.

[18] L. V. Thekkekara and B. Cai. “Significant efficiency enhancement in thin film solar cells using laser beam-induced graphene transparent conductive electrodes”. _eprint: Applied. 2018.

[19] F. D. Torrisi and A. Bemporad. “Discrete-time hybrid modeling and verification”. In: *Proceedings of the 40th IEEE Conference on Decision and Control (Cat.* 3, vol. 3: No. 01CH37228), 2001, pp. 2899–2904.

[20] N. Vasilache, C. Bastoul, and A. Cohen. “Polyhedral code generation in the real world”. In: *Compiler Construction: 15th International Conference, CC 2006, Held as Part of the Joint European Conferences on Theory and Practice of Software, ETAPS 2006, Vienna, Austria, March 30-31, 2006*. Heidelberg: Springer, 2006, pp. 185–201.

[21] Sven Verdoolaege. “isl: An Integer Set Library for the Polyhedral Model”. In: *Mathematical Software – ICMS 2010*. Ed. by David Hutchison et al. Vol. 6327. Series Title: Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 299–302. ISBN: 978-3-642-15581-9 978-3-642-15582-6. DOI: 10.1007/978-3-642-15582-6_49. URL: http://link.springer.com/10.1007/978-3-642-15582-6_49 (visited on 07/10/2024).

[22] Sven Verdoolaege et al. “Polyhedral parallel code generation for CUDA”. en. In: *ACM Trans. Archit. Code Optim.* 9.4 (Jan. 2013), pp. 1–23. ISSN: 1544-3566, 1544-3973. DOI: 10.1145/2400682.2400713. URL: https://dl.acm.org/doi/10.1145/2400682.2400713 (visited on 07/10/2024).

[23] Weichuang Zhang et al. “An Optimizing Framework on MLIR for Efficient FPGA-based Accelerator Generation”. In: *2024 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. Edinburgh, United Kingdom: IEEE, Mar. 2024, pp. 75–90. ISBN: 9798350393132. DOI: 10.1109/HPCA57654.2024.00017. URL: https://ieeexplore.ieee.org/document/10476481/ (visited on 09/20/2024).

[24] Y. Zhang and J. Yang. “Optimizing I/O for Big Array Analytics”. _eprint: abs/1204.6081. 2012.

[25] Ruizhe Zhao et al. “POLSCA: Polyhedral High-Level Synthesis with Compiler Transformations”. In: *2022 32nd International Conference on Field-Programmable Logic and Applications (FPL)*. Belfast, United Kingdom: IEEE, Aug. 2022, pp. 235–242. ISBN: 978-1-66547-390-3. DOI: 10.1109/FPL57034.2022.00044. URL: https://ieeexplore.ieee.org/document/10035220/ (visited on 09/20/2024).

SACI Abdallah (photo not available) is a researcher in the Computer Science Department at Batna 2 University in Algeria. He received his master’s degree from Batna University in Algeria, with a specialization manufacturing scheduling. Currently, he holds the position of assistant professor. His research interests primarily revolve around code generation, with a specialization in Polyhedral model.

SEGHIR Rachid (photo not available) is a full professor at the university of Batna 2, Algeria. He received his engineering degree from the university of Batna in 2000, and DEA and PhD degrees in computer science from the university of Strasbourg, France in 2002 and 2006 respectively. His research interests include symbolic counting problems, program optimization and parallelization, polyhedral model, and parallel computing. He is also involved in some other research areas, such as applied mathematics and natural-inspired optimization.

Arabic Text Summarization using transformer-based architectures

Karim Morsi[*]
Faculty of computer and information science, Ain-Shams University
Fatma najib[†]
Faculty of computer and information science, Ain-Shams University
Wedad Hussein[‡]
Faculty of computer and information science, Ain-Shams University
Rasha Ismail[§]
Faculty of computer and information science, Ain-Shams University

Abstract

Text summarizing is one of the most challenging tasks in natural language processing (NLP). This task is addressed in a large number of research projects and papers in the literature, but most of them focused on English language. Few studies are dealing with the complex Arabic language. Pre-trained Transformer-based language models have shown remarkable efficacy in addressing problems associated with text generation and natural language processing in recent times. However, there has not been much research on applying these models to Arabic text production. This study focuses on the implementation and fine-tuning pre-trained transformer-based language model structures for Arabic abstractive summarization, including AraBERT, mBERT models, and AraT5. We applied mBERT and AraBERT in the context of text summarization using a BERT2BERT-based encoder-decoder model. ROUGE measurements and manual human evaluation have been used to test the suggested models. Our models are trained and tested using XL-Sum Dataset of 46897 high-quality text-summary pairs. Their performance on out-of-domain data was also compared. We found that AraT5 outperforms AraBERT and mBERT Models, suggesting that a pre-trained Transformer with encoder-decoder functionality is more suited for text summarization. Moreover, AraT5 achieve high performance on out-of-domain dataset and received higher accuracy ratings in human evaluations compared to other models.

Key Words:Arabic natural language processing; Abstractive text Summarization; machine learning; Deep learning; Transfer learning models.

1 Introduction

It becomes more difficult to quickly and accurately extract important information from texts due to the massive volume of digital text data generated every day [1]. Moreover, automatic text summarizing is important for many applications. It improves the process of retrieving important data from digital documents through the use of advanced filtering techniques, making it easier to find embedded knowledge in these materials. Additionally, this technology helps manage the enormous amount of textual material that is accessible. Document summarizing helps to overcome the challenges caused by the huge amount of information on the Internet by reducing, organizing, and retrieving information as needed [2]. Additionally, there are several useful applications for text summarizing, such as compressing articles for online publications, optimizing search engine rankings, and making theses and research papers easier to understand. It is also helpful in creating systems for organizing and screening information sources so that only relevant information is taken out of them. Because of its flexibility, text summarization is a useful technique for increasing productivity in a variety of fields and speeding information access.

The extractive and abstractive approaches constitute the two primary types of automated text summarization [3]. The final sentences generated by summaries that only use content that has been extracted contain words or phrases that were taken from the original text. This method is called extractive summary [4], whereas abstractive summarization uses linguistic approaches to comprehend the text and compressed its essential concepts [3]. Depending on the particular needs of the assignment, these strategies are applied in different applications and accommodate alternative methodologies for producing summaries.

It's clear that the field of text summarization has focused mainly on the English language, but dealing with the Arabic language's complexity presents significant challenges. Arabic has special difficulties because of its complex morphology, diglossia, and diversity of dialects. Compared to English, automatic summarization in Arabic is a more difficult to

[*] Faculty of computer and information science, Ain-Shams University, Cario, Egypt. Email: karim.mohamed.fcis@cis.asu.edu.eg.

[†] Faculty of computer and information science, Ain-Shams University, Cario, Egypt. Email: fatma_mohamed@cis.asu.edu.eg.

[‡] Faculty of computer and information science, Ain-Shams University, Cario, Egypt. Email: wedad.hussein@cis.asu.edu.eg.

[§] Faculty of computer and information science, Ain-Shams University, Cario, Egypt. Email: rashaismail@cis.asu.edu.eg.

apply because of these linguistic features. Moreover, the vast majority of text summarization systems currently in use, including those for Arabic, depend on extractive summary strategies. In particular in the Arabic context, abstractive summarization is less frequent. However, the reality that Arabic is the official language of 22 countries and is spoken by over 300 million people shows how important it is to address these issues in Arabic text summarization [2, 5]. Effective Arabic summarization systems are becoming more and more necessary in order to support the efficient processing and retrieval of information in the Arab-speaking world. The significance of developing text summarizing methods specific to the Arabic language's varied requirements and distinctive linguistic characteristics is being acknowledged by researchers and developers [6].

Text translation [7], sentiment analysis [8], text summarization, and other critical tasks have recently shown significant improvements caused by deep learning techniques [9]. Moreover, using large datasets to enhance performance is a key component of deep neural network applications [10]. The encoder-decoder model's sequence-to-sequence structure serves as the foundation for the new text summarizing techniques. The encoder and decoder are the two components of this paradigm. The encoder changes the hidden states in accordance with each new token it gets from the input sequence at each step. Regardless of the length of the input, the encoder creates a context vector representing the input sequence when it reaches the final token in the sequence. The context vector is the final hidden state to be established before the decoder. The decoder is started with ⟨SOS⟩ token, and context vector from the encoder as a first hidden state is used to start it. The decoder is taught to generate a new sequence with a predetermined length. By providing the previously created word, the device creates a new word from the vocabulary each time [10, 11]. As seen in Figure 1, the start token ⟨SOS⟩ [12] is supplied to the decoder along with the encoder's final hidden state. Numerous NLP applications, including machine translation and text summarization, have made use of this approach. A sentence in specific language is the input sequence for machine translation, while the output sequence is the same statement in a different language. In contrast, the document that has to be summarized is the input sequence in text summarization, and the summary itself is the output sequence [12, 13].

The traditional sequence-to-sequence (seq2seq) model faces a challenge in summarizing lengthy input sequences by compressing them into a single fixed-size ”context vector”. This method often struggles to capture all essential details, especially in longer sequences.

attention mechanism allows the model to selectively highlight important parts of the input sequence when generating each part of the output sequence as shown in Figure 2, By dynamically assigning significance scores to different segments of the input sequence during decoding, applying encoder hidden states, the attention mechanism enables the model to better understand the relevance of individual elements. This adaptability helps in focusing on important information within longer input sequences, significantly improving the model's accuracy and performance in tasks like machine translation and summarization [13].

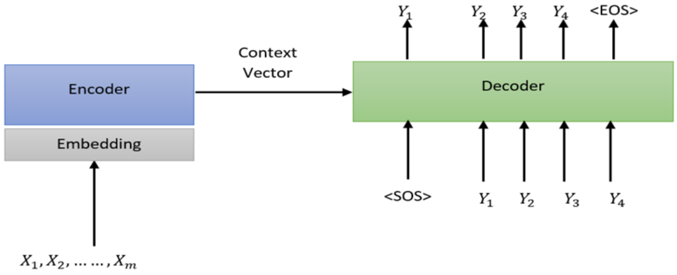


Figure 1: Sequence-to-sequence model.

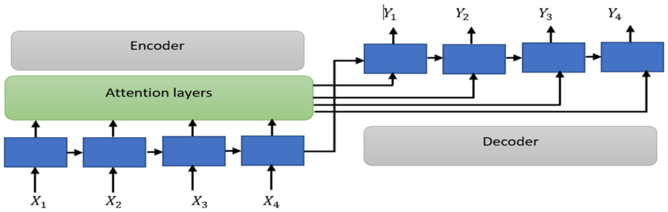


Figure 2: Sequence-to-sequence with attention.

Attention processes come in two varieties, such as local and global attention. The context vector's derivation method determines how they vary from one another. The attended context vector in global attention is derived from all of the encoder's hidden states, whereas in local attention it is derived from a limited number of encoder hidden states [13].

The Transformer model architecture was introduced in recent years, which has allowed for performance benefits over RNN-based designs [14]. Additionally, a fresh approach known as transfer learning has surfaced and grown rapidly to take the lead in deep learning model training and application. Using a self-supervised training goal, the model is pre-trained on a vast quantity of data in the first phase of this technique. Using a supervised data set, the model is then fine-tuned on a downstream job in the second phase [15].

We used the XL-Sum dataset to fine-tune the pre-trained models mBERT, AraBERT, and AraT5 in this study. Additionally, mBERT and AraBERT have been fine-tuned using the BERT2BERT architecture. ROUGE measures were employed together with manual human evaluation to evaluate the performance of the suggested models. Their results on a test set outside of their field of study were also compared.

The rest of the paper is organized as follows. Section 2 discusses the related studies to our work. The methodology of the proposed work is presented in section 3. Section 4 presents the setup of our experiments. Results are discussed in section 5. Section 6 introduces the conclusion of the paper.

2 Related works

This section presents recent studies on Arabic abstractive text summarization. We will concentrate on the studies that used the Transformer architecture in addition to the ones that used the RNN-based sequence-to-sequence model.

It has been shown by several recent studies that transfer learning produces state-of-the-art outcomes on nearly all NLP tasks [16]. This shows that the skills acquired through neutral unsupervised learning may be effectively applied to challenges that come after. The NLP community has recently seen an increase in the use of big pre-trained models that have used this methodology as a result of its success [17, 18].

The Transformer-based pre-trained models and RNN-based architecture have not been widely used in Arabic-language works [19].[20] have improved mBERT [21], mBART-50 [22], and AraBERT [23] for cross-lingual Arabic abstractive text summarization, and found that AraBERT produces the lowest result of any of their other suggested models. AraT5 has also been improved by [24] for Arabic abstractive summarization multi-sentence.

In [19] they provided a comprehensive comparative analysis between RNN-based and Transformer-based architectures, specifically, mBERT, AraBERT, AraGPT2, and AraT5, which are well-known for their ability to understand and produce Arabic text for tasks requiring abstractive summarization. Their paper involved a sizable Arabic summarization dataset, contains 84,764 high-quality text-summary pairs, serving as both training and evaluation data. To combat potential under-fitting, an additional dataset of 280,000 examples was incorporated, leading to the improvement and enhancement of models, denoted as ”Seq2Seq-LSTM+” and ”Transformer+”. Notably, these ”+” variants, trained on the expanded dataset, shown improved performance. Achieving F-scores of 33.04 for Seq2Seq-LSTM, 32.12 for Transformer, 37.57 for Seq2Seq-LSTM+, and notably higher scores of 39.61 for Transformer+, 42.96 for mBERT2-mBERT, 40.48 for AraGPT2, 44.02 for BERT2BERT, and the highest of 46.87 for AraT5 using ROUGE-L.

A hybrid approach was developed for Arabic summarization by combining a transformer-based model with a Modified Sequence-To-Sequence (MSTS) framework [25]. (MSTS) model involves the incorporation of three encoder layers, specifically input layer, sentence layer, and named entity recognition layers, aimed at improving the summarization process. They used global attention mechanism and AraVec for embedding and building a new dictionary to cover the word that not included in AraVec. This innovative strategy involves enhancing the MSTS model. By selectively choosing and rearranging text fragments, the model generates extractive summaries. Subsequently, the transformer-based mechanism refines these extractive summaries, transforming them into abstractive summaries. The HASD (Arabic Summarization Dataset) was introduced as a novel benchmark dataset and the existing extractive EASC benchmark was modified by incorporating abstractive summaries into each text. To evaluate the quality of abstractive summaries, they proposed a new evaluation metric termed the Arabic-ROUGE measure. This metric evaluates vocabulary and structural similarity of abstractive summaries, emphasizing their coherence and linguistic essence.

In contrast, however, the study presented by [26] proposed abstractive summarization system used a sequence-to-sequence (seq2seq) model enhanced with different recurrent neural network (RNN) architectures, which are Gated Recurrent Units (GRU), Long Short-Term Memory (LSTM), and Bidirectional LSTM (BiLSTM). Both the encoder and decoder components integrated global attention mechanisms, allowing the model to focus on relevant parts of the input during encoding and decoding. To enhance the understanding of Arabic words and achieve improved performance, the AraBERT preprocessing stage was incorporated into the model pipeline. Furthermore, a comparative study was conducted between two Word2Vec models, skip-gram and continuous bag of words (CBOW). The study showed that employing a Bidirectional LSTM (BiLSTM) architecture, consisting of three hidden layers, and integrating AraBERT preprocessing led to superior performance results. This finding suggests the advantage of the BiLSTM architecture in conjunction with AraBERT preprocessing for enhancing the abstractive summarization of Arabic text. They used the Arabic Headline Summary (AHS) and the Arabic Mogalad_Ndeef (AMN) datasets.

An extractive summarization system was introduced in [27] .The initial phase involved organizing an Arabic text into a graph format, where sentences act as nodes connected by edges representing similarity. Using cosine similarity, sentences exceeding a set threshold were linked, creating a highly interconnected graph. Employing the PageRank algorithm on this weighted graph assigned salience scores to each sentence, determining their significance within the network. Iteratively computed based on edge weights and damping factor, these scores identified sentences that are relevant and strongly connected. Subsequently, sentences were ranked according to their salience scores, organizing them in order of importance. This process ensured that sentences with stronger relationships and relevance stand out. Triangles within the graph were identified using De-Morgan laws, aiding in constructing a reduced graph that captured the essential elements of the text. They tested their model using EASC dataset.

An advanced text summarization model was proposed in [28] based on a sequence-to-sequence RNN architecture, specifically using LSTM units to reduce the vanishing gradient problem. Diverging from a single-layer encoder, it employed a three-layered multilayer encoder. One layer captures input text, another grasps text keywords, and the third identifies text name entities, all facilitated by word embeddings. The hidden states of the three encoder layers consist of bidirectional LSTM units. The decoder, a singular unidirectional LSTM layer, receives training input through attention to previous summary words and decoder hidden states. In testing, it depends on the previous decoder output and hidden states, initialized with "<SOS>" and the context vector. Employing global attention mechanisms enhanced prediction accuracy by adding important source text insights into the context vector. A dataset comprises 79,965 documents was used. This dataset sourced from news sites like Aljazeera, National Interest, and Financial Times along with 69,024 documents from SANAD_SUBSET, categorized into medical, finance, sports, religion, culture, politics, and technology.

In [29] they introduced finetuning the AraBART model, based on BART Base architecture, comprises 6 encoder and 6 decoders layers with 768 hidden dimensions, totaling 139M parameters. It incorporates an extra layer-normalization for stable training at FP16 precision and uses sentence piece for a 50K token vocabulary and 99.99% character coverage of the training data. This model was evaluated on datasets like Arabic Gigaword subsets and XL-Sum, it covered various abstractiveness levels in news articles and includes tasks for summary and title generation. The comparison was against baselines, AraBART is compared to C2C (a BERT2BERT-based seq2seq model), mBART25 (pretrained on 25 languages, fine-tuned for Arabic), and mT5base. They found AraBART consistently surpasses C2C and mBART25 across various datasets, Its superiority, based model. In this work, we use various Transformer-based models abstractiveness. Additionally, AraBART outperforms mT5 on the multilingual setup for XL-Sum.

In [30] an automatic and extractive method was proposed for single-document summarization in the Arabic language. The proposed method aims to create informative summaries by evaluating each sentence's importance based on a combination of statistical and semantic features like (Key-Phrases, Sentence location, Similarity with title, Sentence centrality, Sentence length, Cue words, Positive key-words, Sentence inclusion of numerical data, Occurrence of Non-essential Information). In the score-based method, important sentences were extracted based on the total scores that are assigned to them. In the machine learning approach, the extractive summarization process was modeled as a binary classification problem Then, a binary (Yes/No) classifier was trained based on a set of training documents.

In [31] they introduced SemG-TS, an innovative Arabic abstractive summarization technique based on semantic graph embeddings and a deep neural network. SemG-TS transforms text into a semantic graph, capitalizing on Arabic language nuances, followed by SemanticGraph2Vec graph embedding. The deep learning model based on a sequence-to-sequence architecture used in summarization includes LSTM in the Encoder and LSTM Basic Decoder. Using AlJazeera.net data comprising 16,770 paragraphs averaging 204 words each, SemG-TS exceeds two word2vec versions (trained and random-based) across ROUGE metrics. It achieved a 15.8% precision improvement, 29.5% in recall, and 21.4% in F-measure over the best word2vec (random-based). In human evaluation, SemG-TS shows superior relevancy, similarity, readability, and overall satisfaction compared to word2vec. The F-score for ROUGE stands at 0.047.

Previous Arabic text summarization studies focused on RNN-based model. In this work, we use various Transformer-based models. mBERT,AraBERT and AraT5 pre-trained language models were used. A high-quality dataset was used to compare the performance achieved by these models.

3 Methodology

First, as seen in figure 3, we explain the dataset preparation process that was employed in this part. Next, we go into the model architecture and training details of the several models that have been trained for the Arabic abstractive text summarizing work.

3.1 Dataset

We choose the XL-Sum dataset [32], which is appropriate for the abstractive summarization of a single document. One million texts with clear summary are included. With the use of well-designed algorithms, this dataset was generated from news articles on the BBC website. The 44 languages in the XL-Sum dataset have available ranging from low to high, with many not having public access.

3.2 Data Preprocessing

Arabic has greater difficulties than some other languages, like English, because it is a morphologically rich language. The inconsistent use of diacritical marks (Tashkil) and the omission of Hamza in Arabic texts provide difficulties for the processing of Arabic text. Modern Standard Arabic (MSA) is the Arabic language used most commonly in academic work, news, and literature, it often omits tashkil. Because the Arabic text missing the diacritical representations necessary to change a word's meaning, this omission increases ambiguity. Additionally, various dialects are spoken in Various Middle Eastern areas, each with significant differences. This diversity in dialects further complicates language processing and understanding in Arabic text analysis and summarization tasks. Each of these difficulties must be taken into consideration while processing Arabic text. The AraBERTv1-base Model [23] was used for text preprocessing. It is a powerful language model that facilitates Arabic text processing and analysis across a wide range of NLP tasks.

3.3 Data Tokenization

Tokenization is a key component of natural language processing research because it covers the gap between unprocessed textual data and the numerical input needed to build machine learning models. Tokenizer classes were strategically used to enable effective model training and evaluation, as well as efficient data preprocessing and easy integration with the corresponding models. Tokenization process was

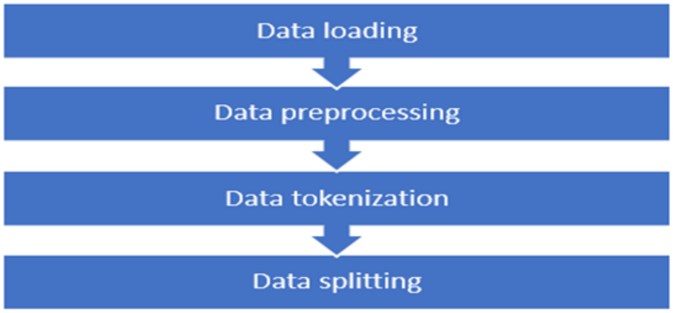


Figure 3: The steps for preparing the data.

used for improving the efficiency and adaptability of our trained models. The Hugging Face Transformers library's AutoTokenizer class was used for improving the performance of AraT5 model. Furthermore, BertTokenizer was used for optimizing the BERT2BERT and mBERT2mBERT models. In our experiments, the AutoTokenizer and BertTokenizer were used for truncating the input sequence to 512 token and padding the short sequence to it. For text summary, text was truncated to100 token and padding the short summary to it.

3.4 Data Splitting

The XL-Sum dataset includes abstractive summaries of 46897 Arabic articles that were created by humans. We split the dataset to 37519 (80% of the records) articles for training, 4689 (10%) articles for validation and 4689 (10%) articles for testing.

3.5 Building an outside domain set

We used an additional sampled dataset from the Arabic Mogalad_Ndeef Dataset (AMN) focused on single-sentence abstractive summarization [33]. Random samples were chosen about 1000 records. However, we did not train the model on this subset of the data.

4 Experiments

The architecture, experimental details, and model training process are described in detail below. Figure 4 presents the architecture of the models used in this research.

4.1 BERT2BERT

AraBERT, an encoder-only Transformer, is the Arabic version of BERT. It accepts an input of length n, called $X_{1:n}$, and generates a contextual representation based on that input, with same length $X^{-}{}_{1:n}$. AraBERT isn't appropriate for text summarization because it requires input and output lengths to match, which presents a limitation when summarizing text, as the length of the original text and the summary may differ. To apply the BERT2BERT encoder-decoder setup, we initialized both the encoder and decoder with AraBERT weights, enabling the utilization of AraBERT for summarization purposes [19]. To build our model, we included AraBERT parameters into the appropriate BERT2BERT layers. The decoder part was significantly modified, but the encoder component matches AraBERT without modifying its settings. In every block of the decoder, we added cross-attention layers, which had weights that were originally randomly distributed across the feed-forward and self-attention layers. Additionally, we converted bidirectional self-attention layers into unidirectional ones so that the decoder only analyzes tokens that have previously been created at each step. After the last decoder block, we added an LM head to allow for the creation of summary tokens. The final layer was initialized in the same way as the embedding layer. We suggested to reduce the total number of trainable parameters by sharing the encoder's weights with the decoder because of their close similarity. During initialization, in decoder blocks only the cross-attention layers were randomly initialized. We trained the model for 5 epochs with a batch size of 2. Before backpropagation, we collected gradients during fine-tuning for 8 steps. We used ”bert-base-arabertv02” as the version of AraBERT.

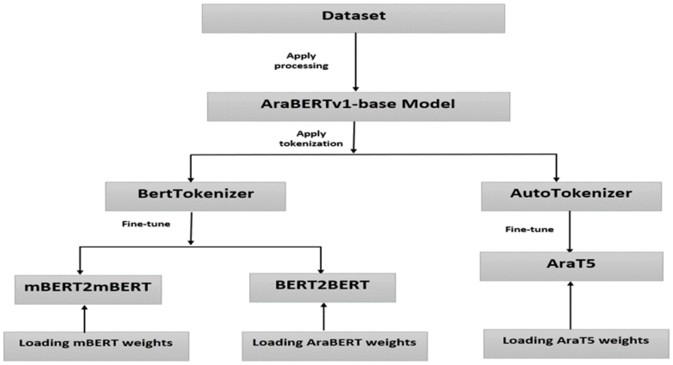


Figure 4: A diagram of the trained models.

4.2 mBERT2mBERT

In our approach, we used the BERT2BERT encoder-decoder architecture with mBERT initialization, a model pre-trained on a diverse corpus covering 104 languages, including Arabic sourced from Wikipedia text. During training, we implemented a strategy of accumulating gradients over 8 steps before proceeding with back-propagation, a technique aimed at saving the training process and improving convergence. Fine-tuning of the model was conducted over 5 epochs, with a relatively small batch size of 2, chosen to balance computational efficiency and model performance. To accommodate the characteristics of text summarization tasks, we restricted the length of both input sequences to 512 tokens and summaries to 100 tokens during fine-tuning. This limitation helps in managing computational resources effectively while ensuring that the model can capture essential information for summarization within the specified constraints. Overall, these training and fine-tuning strategies are designed to optimize the BERT2BERT architecture for Arabic text summarization, aiming to achieve robust performance across a range of summarization tasks and dataset.

4.3 AraT5

A modified version of the popular T5 model is the AraT5 [34], serves as an encoder-decoder framework that combines various natural language processing tasks within a single text-to-text paradigm. Because of its flexibility, AraT5 can easily handle a wide range of tasks, including text summarization, machine translation, and categorization. Notably, the model utilizes task-specific prefixes appended to input sequences to determine the kind of task, such as" translate English to Arabic" for translation or" summarize:" for summarization. During training, approximately 15% of the tokens in this model's training set are masked in Masked Language Modeling (MLM), with consecutive tokens being masked using a single sentinel token [19]. For our experimentation, we leveraged the AraT5$_{\text{Base}}$ version 8, which underwent training on a diverse dataset comprising both Modern Standard Arabic (MSA) and Twitter data, utilizing the T5$_{\text{Base}}$ architecture. The MSA dataset, totaling 70 GB, was sourced from various Arabic repositories, while the Twitter data encompassed 1.5 billion tweets containing at least three Arabic words, randomly sampled for inclusion. The architecture of the encoder and decoder is identical to that of BERTBase, consisting of 12 layers with 12 attention heads each.

During fine-tuning, the model underwent training for 5 epochs with a batch size of 2. The input sequence length was covered at 512 tokens, with summaries restricted to 100 tokens.

In this work, several huge trained models were used. Firstly, we optimized the multilingual mBERT model, which is commonly used as a baseline in the literature. Next, improvements were implemented to the Arabic pre-trained models AraBERT and AraT5. We made use of the BERT2BERT encoder-decoder architecture to use AraBERT and mBERT for summarizing texts, where the corresponding model weights were used to warm-start the encoder and decoder . For fine-tuning the pre-trained models, we used Adam optimizer (Adaptive Moment Estimation) [35], which is an optimization algorithm commonly used in training deep learning models, including those used in natural language processing (NLP) tasks. It belongs to the family of stochastic gradient descent (SGD) optimization algorithms and is known for its efficiency and effectiveness in a wide range of applications, with a learning rate of 2e-5.

5 Experimental results and Discussion

We fine-tune three transformer models AraT5, AraBERT, and mBERT and apply improvements to better adapt them for the task of Arabic summarization. We then evaluate their performance and compare the results. Text summarization models are usually evaluated automatically using ROUGE metrics in addition to be assessed manually by human experts. While ROUGE metrics quantify overlap between generated and reference summaries, they may not fully capture qualitative aspects like coherence and readability. Manual evaluation supplements this by considering factors such as relevance, coherence, and grammaticality, providing a more detailed understanding of summary quality. By combining automated and manual evaluations, researchers gain a comprehensive view of model performance, enabling model selection and strategies for improvement. This double evaluation strategy provides an in-depth review, allowing for well-informed choices to be made in the development and research of text summarizing.

5.1 Automatic evaluation

We used the ROUGE-1, ROUGE-2, and ROUGE-L measures for automated evaluation. These metrics measure the degree to which produced summaries and reference summaries overlap in terms of unigrams, bigrams, and longest common subsequences. The model's performance in comparison to the reference summary was then measured by computing the accuracy, recall, and F-measure values for each ROUGE metric. The following formulas were used to get values for each ROUGE metric:

$$\mathbf{precision} = \frac{\left|\mathbf{grams_{refrence}} \cap \mathbf{grams_{generated}}\right|}{\mathbf{grams_{generated}}} \tag{1}$$

$$\mathbf{recall} = \frac{\left|\mathbf{grams_{refrence}} \cap \mathbf{grams_{generated}}\right|}{\mathbf{grams_{generated}}} \tag{2}$$

$$\mathbf{F-measure} = \mathbf{2.0} * \frac{\mathbf{recall} * \mathbf{precision}}{\mathbf{recall} + \mathbf{precision}} \tag{3}$$

The evaluation results, presented in terms of ROUGE F1 scores, were obtained using the rouge Python library. During summary generation, we used Beam Search algorithm which is a heuristic search algorithm commonly used in sequence generation tasks, efficiently explores the search space by maintaining a set of candidate sequences, selecting the most promising candidates at each step based on a predefined scoring criterion, and pruning less likely paths to focus on high-quality outputs [36]. The beam search algorithm was employed with a beam size of 3.

In our experimental setup, we initially trained and evaluated models using a subset of the data, consisting of 10000 records for training and 1200 records for evaluation and testing, and then we used the full dataset. Figure 5 provides ROUGE F1 scores evaluation on test set. Arat5 performed better than other models on the test set. Figure 6 shows the ROUGE F1 scores evaluation on validation set. BERT2BERT which was initialized with AraBERT weights had high performance on the validation set. Additionally, within our model comparison, AraT5 outperformed mBERT2mBERT on the validation set, showing

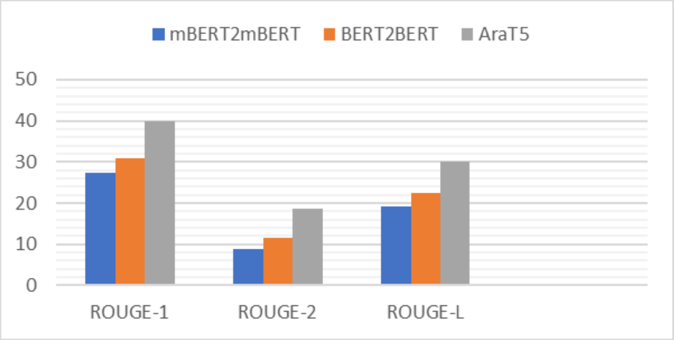


Figure 5: Test set scores based on the ROUGE F1 evaluation.

even more the differential efficiency of different architectures in text summarizing tasks. Additionally, we evaluated our models using the 1000-record outside domain set. as shown in figure 7, and found that AraT5 achieved the highest score compared to other models. As a result, we used AraT5 model to compare our work with other comparing studies.

5.2 Manual evaluation

As was previously mentioned, a thorough assessment of the quality and readability of produced summaries could not be possible if one only relies on ROUGE indicators [19]. Therefore, we handled human evaluations. Five fluent Arabic speakers were tasked with rating each summary on a scale of one to five based on two criteria: (1) readability, which measures grammatical accuracy and sentence structure; and (2) quality, which assesses how well the summary conveys the main ideas of the original text. We chose 20 cases at random from the test set. The human evaluation criteria are listed in Table 1. Next, we determined the reported scores' mean. The results of the manual human examination are shown in Figure 8.

The two models with the highest scores on the two measures were AraT5 and BERT2BERT, with AraT5 almost outperforming BERT2BERT. On the other hand, mBERT2mBERT scored significantly lower for both quality and readability measures. With the use of our models, we developed summaries for the two articles, which are shown in Figure 9 and 10.

5.3 Comparison with previous studies

Table 2 presents a comparison of ROUGE F1 evaluations on the XL-Sum dataset between our AraT5 model and four types of state-of-the-art baseline results [29]. The initial baseline, named C2C, is a monolingual sequence-to-sequence model [37], which is based on BERT2BERT. While the cross-attention weights are initialized at random, the encoder and decoder are initialized using CAMELBERT weights [38]. A total of 246M parameters represent C2C. The multilingual BART model mBART25 [39], pretrained on 25 languages, including Arabic, is the second baseline. mBART25 has shown successful in monolingual generative tasks like abstractive summarization, although it was initially pre-trained for neural machine translation [40]. mBART25 has 610M parameters overall. Third model is called mT5$_{\text{base}}$ model. Finally, AraBART, the fourth model, performs better than the others. As noticed our model outperforms the four compared state-of-the-art models.

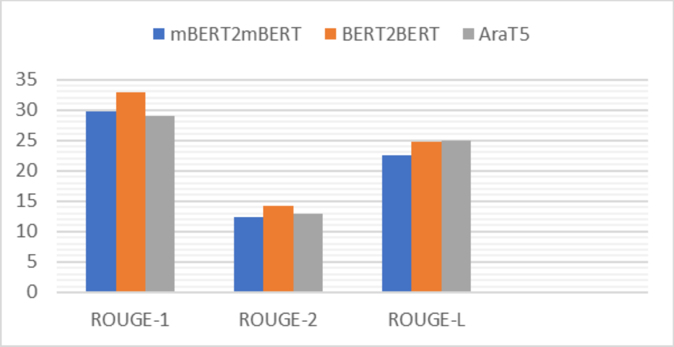


Figure 6: Validation set scores based on the ROUGE F1 evaluation.

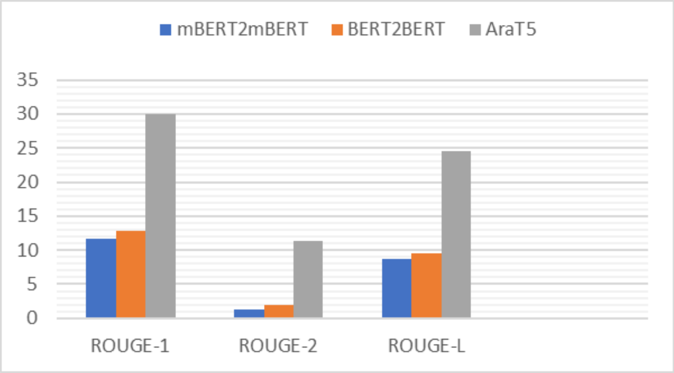


Figure 7: ROUGE F1 scores of the out-of-domain set.

5.4 Discussion and findings

ROUGE metrics indicate that fine-tuning the AraT5 model resulted in a performance increase of approximately 15.56%. In contrast, the performance of AraBERT and multilingual BERT decreased by 10.52% and 20.67%, respectively. When summarizing data from outside the domain, AraT5 performs better than AraBERT Model. Interestingly, the BERT2BERT-based model initialized using multilingual BERT shows poor performance when evaluated manually. In comparison to the models that have been recommended, AraT5 and AraBERT model consistently generate highest score summaries in terms of readability and quality, as assessed by human evaluation. Furthermore, it's observed that increasing the training data enhances the model's accuracy.

5.5 Limitations

While our research focused mostly on producing one-sentence summaries from news sources, there is still a need

Table 1: Readability and quality measures for the manual human evaluation.

Score	Quality	Readability
1	The output is irrelevant.	Inaccurate / Hard to read.
2	Key concepts are partially conveyed.	A little understood.
3	Key concepts are moderately conveyed.	Understandable in poor Arabic.
4	Key concepts are largely conveyed.	Understandable in acceptable Arabic.
5	Key concepts are completely conveyed.	Understandable in fluent Arabic.

Table 2: A comparison of different state-of-the-art models' ROUGE F1 results.

Model	ROUGE-1	ROUGE-2	ROUGE-L
C2C	26.9	8.7	23.1
mBART25	32.1	12.5	27.6
mT5base	32.8	12.7	28.7
AraBART	34.5	14.6	30.5
AraT5 (Ours)	39.78	18.77	30.21

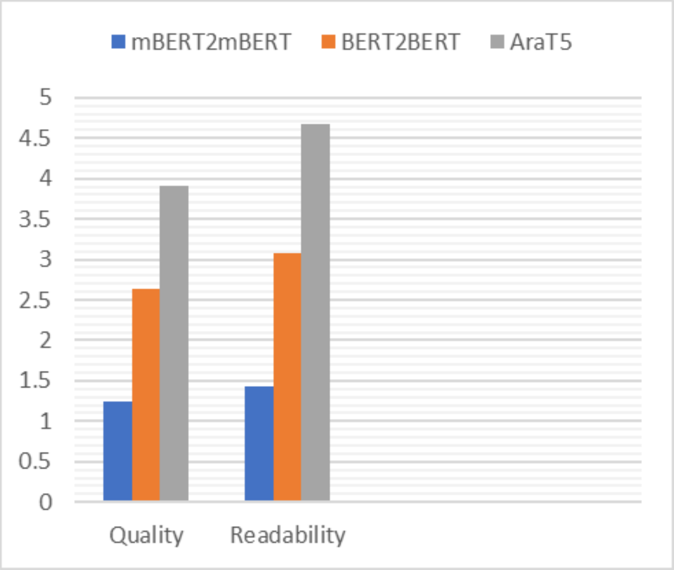


Figure 8: The manual human evaluation scores.

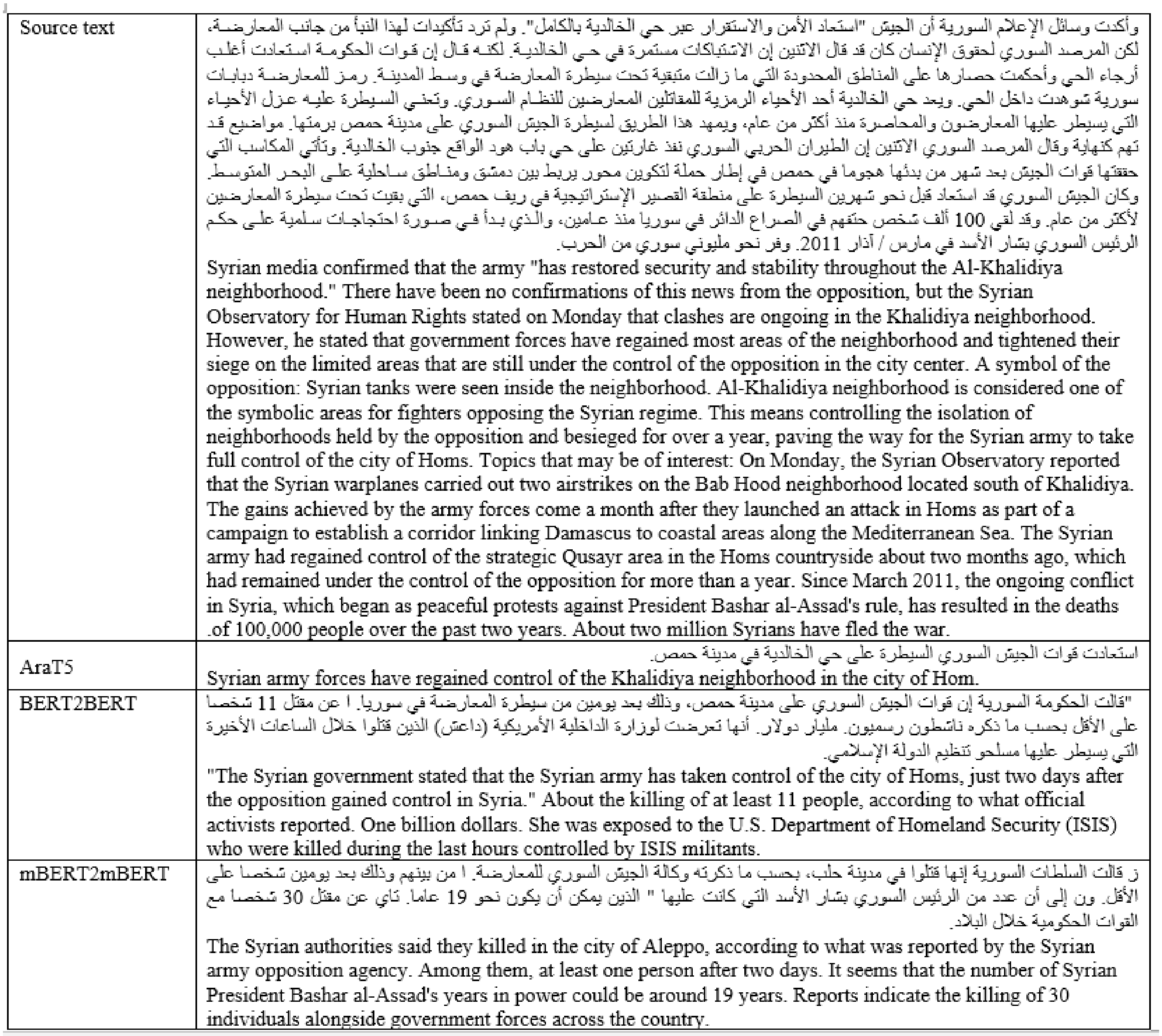
for further study into producing multi-sentence summaries and extending the application to other types of text sources. Further research may examine the complexity involved in summarizing information across several phrases, as well as the Arabic dialects by various text. Our models have been modified and specially designed for text summary of news. As such, we expect that without extra training data customized for particular summarization task, their performance might not be directly comparable to other models. We recognize that on sometimes, our models provide inaccurate, invalid, and grammatical outputs that could lead to general users being confused.

6 Conclusion and Future Works

For the purpose of this work, we used several pre-trained language models, such as AraT5, AraBERT, and mBERT, to summarize Arabic abstractive text. Additionally, to use encoder-only Transformer models, we used an encoder-decoder architecture based on BERT2BERT. Both manual assessment and ROUGE metrics were used to evaluate these models, and an out-of-domain dataset was used for testing. Our results show that pretrained language models perform well in Arabic text summarization tasks. According to the automated evaluation, AraT5 outperforms other models in our test set, but AraBERT outperforms other models in the validation set during training. In addition, human evaluation shows that AraT5 achieves high accuracy in terms of readability and quality. The summary generated by AraT5 is highly comparable to the reference summary. Furthermore, AraT5 outperforms other models using out-of-domain datasets. Results confirmed that the modified AraT5 performing better than other models. For future research directions, we recommend focusing on multi-sentence summarization, emphasizing grammatical correctness, understanding dialects, and incorporating semantic meaning into automatic evaluation processes. In order to improve the model's performance across multiple domains, we propose training it on multi-domain datasets. By using this method, the model's efficiency and adaptability in summarization tasks covering many topics and domains can be enhanced. While we recommend exploring fine-tuning strategies, hybrid models, and additional linguistic resources to further optimize summarization performance.

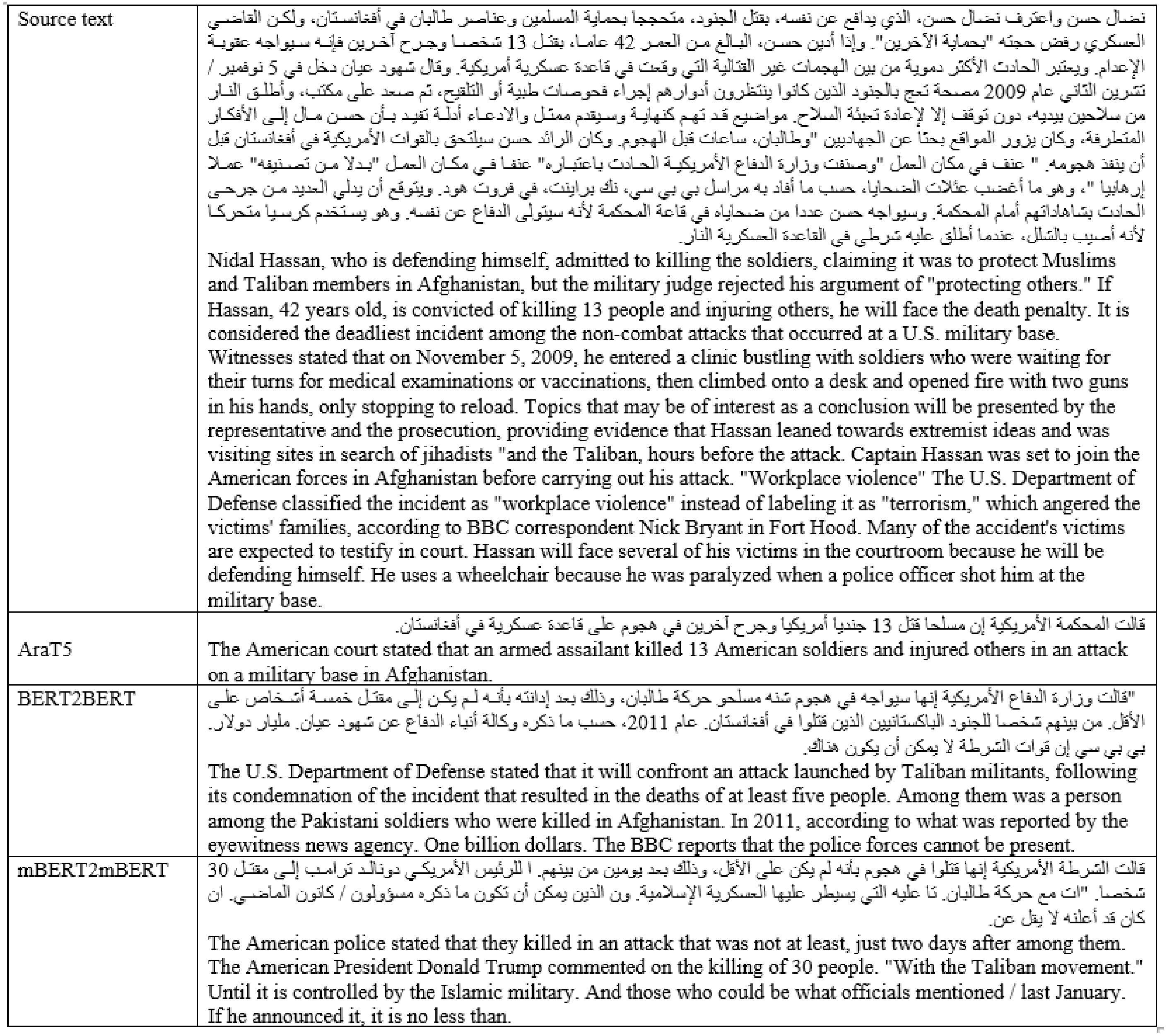
7 Future Work

In future research, we propose focusing on multi-sentence summarization with an emphasis on grammatical correctness, dialect comprehension, and the integration of semantic meaning into automatic evaluation metrics. Expanding the training data to include multi-domain datasets can further enhance the model's adaptability and performance across various topics and domains. To optimize summarization quality, future studies should explore advanced fine-tuning strategies, hybrid modeling approaches, and the incorporation of additional linguistic resources. Investigating the impact of diverse embedding techniques, reinforcement learning-based training, and transformer-based architectures could also contribute to



Source text	وأكدت وسائل الإعلام السورية أن الجيش "استعاد الأمن والاستقرار عبر حي الخالدية بالكامل". ولم ترد تأكيدات لهذا النبأ من جانب المعارضة، لكن المرصد السوري لحقوق الإنسان كان قد قال الاثنين إن الاشتباكات مستمرة في حي الخالدية. لكنه قال إن قوات الحكومة استعادت أغلب أرجاء الحي وأحكمت حصارها على المناطق المحدودة التي ما زالت متبقية تحت سيطرة المعارضة في وسط المدينة. رمز للمعارضة دبابات سورية شوهدت داخل الحي. ويعد حي الخالدية أحد الأحياء الرمزية للمقاتلين المعارضين للنظام السوري. وتعني السيطرة عليه عزل الأحياء التي يسيطر عليها المعارضون والمحاصرة منذ أكثر من عام، ويمهد هذا الطريق لسيطرة الجيش السوري على مدينة حمص برمتها. مواضيع قد تهم كنهاية وقال المرصد السوري الاثنين إن الطيران الحربي السوري نفذ غارتين على حي باب هود الواقع جنوب الخالدية. وتأتي المكاسب التي حققتها قوات الجيش بعد شهر من بدئها هجوما في حمص في إطار حملة لتكوين محور يربط بين دمشق ومناطق ساحلية على البحر المتوسط. وكان الجيش السوري قد استعاد قبل نحو شهرين السيطرة على منطقة القصير الإستراتيجية في ريف حمص، التي بقيت تحت سيطرة المعارضين لأكثر من عام. وقد لقي 100 ألف شخص حتفهم في الصراع الدائر في سوريا منذ عامين، والذي بدأ في صورة احتجاجات سلمية على حكم الرئيس السوري بشار الأسد في مارس / آذار 2011. وفر نحو مليوني سوري من الحرب. Syrian media confirmed that the army "has restored security and stability throughout the Al-Khalidiya neighborhood." There have been no confirmations of this news from the opposition, but the Syrian Observatory for Human Rights stated on Monday that clashes are ongoing in the Khalidiya neighborhood. However, he stated that government forces have regained most areas of the neighborhood and tightened their siege on the limited areas that are still under the control of the opposition in the city center. A symbol of the opposition: Syrian tanks were seen inside the neighborhood. Al-Khalidiya neighborhood is considered one of the symbolic areas for fighters opposing the Syrian regime. This means controlling the isolation of neighborhoods held by the opposition and besieged for over a year, paving the way for the Syrian army to take full control of the city of Homs. Topics that may be of interest: On Monday, the Syrian Observatory reported that the Syrian warplanes carried out two airstrikes on the Bab Hood neighborhood located south of Khalidiya. The gains achieved by the army forces come a month after they launched an attack in Homs as part of a campaign to establish a corridor linking Damascus to coastal areas along the Mediterranean Sea. The Syrian army had regained control of the strategic Qusayr area in the Homs countryside about two months ago, which had remained under the control of the opposition for more than a year. Since March 2011, the ongoing conflict in Syria, which began as peaceful protests against President Bashar al-Assad's rule, has resulted in the deaths .of 100,000 people over the past two years. About two million Syrians have fled the war.
AraT5	استعادت قوات الجيش السوري السيطرة على حي الخالدية في مدينة حمص. Syrian army forces have regained control of the Khalidiya neighborhood in the city of Hom.
BERT2BERT	"قالت الحكومة السورية إن قوات الجيش السوري على مدينة حمص، وذلك بعد يومين من سيطرة المعارضة في سوريا. ا عن مقتل 11 شخصا على الأقل بحسب ما ذكره ناشطون رسميون. مليار دولار. أنها تعرضت لوزارة الداخلية الأمريكية (داعش) الذين قتلوا خلال الساعات الأخيرة التي يسيطر عليها مسلحو تنظيم الدولة الإسلامي. "The Syrian government stated that the Syrian army has taken control of the city of Homs, just two days after the opposition gained control in Syria." About the killing of at least 11 people, according to what official activists reported. One billion dollars. She was exposed to the U.S. Department of Homeland Security (ISIS) who were killed during the last hours controlled by ISIS militants.
mBERT2mBERT	ز قالت السلطات السورية إنها قتلوا في مدينة حلب، بحسب ما ذكرته وكالة الجيش السوري للمعارضة. ا من بينهم وذلك بعد يومين شخصا على الأقل. ون إلى أن عدد من الرئيس السوري بشار الأسد التي كانت عليها " الذين يمكن أن يكون نحو 19 عاما. تاي عن مقتل 30 شخصا مع القوات الحكومية خلال البلاد. The Syrian authorities said they killed in the city of Aleppo, according to what was reported by the Syrian army opposition agency. Among them, at least one person after two days. It seems that the number of Syrian President Bashar al-Assad's years in power could be around 19 years. Reports indicate the killing of 30 individuals alongside government forces across the country.

Figure 9: A sample text that our models have summarized.



Source text	نضال حسن واعترف نضال حسن، الذي يدافع عن نفسه، بقتل الجنود، متحججا بحماية المسلمين وعناصر طالبان في أفغانستان، ولكن القاضي العسكري رفض حجته "بحماية الآخرين". وإذا أدين حسن، البالغ من العمر 42 عاما، بقتل 13 شخصا وجرح آخرين فإنه سيواجه عقوبة الإعدام. ويعتبر الحادث الأكثر دموية من بين الهجمات غير القتالية التي وقعت في قاعدة عسكرية أمريكية. وقال شهود عيان دخل في 5 نوفمبر / تشرين الثاني عام 2009 مصحة تعج بالجنود الذين كانوا ينتظرون أدوارهم إجراء فحوصات طبية أو التلقيح، ثم صعد على مكتب، وأطلق النار من سلاحين بيديه، دون توقف إلا لإعادة تعبئة السلاح. مواضيع قد تهم كنهاية وسيقدم ممثل والادعاء أدلة تفيد بأن حسن مال إلى الأفكار المتطرفة، وكان يزور المواقع بحثا عن الجهاديين "وطالبان، ساعات قبل الهجوم. وكان الرائد حسن سيلتحق بالقوات الأمريكية في أفغانستان قبل أن ينفذ هجومه. " عنف في مكان العمل "وصنفت وزارة الدفاع الأمريكية الحادث باعتباره" عنفا في مكان العمل "بدلا من تصنيفه" عملا إرهابيا "، وهو ما أغضب عائلات الضحايا، حسب ما أفاد به مراسل بي بي سي، نك براينت، في فروت هود. ويتوقع أن يدلي العديد من جرحى الحادث بشاهاداتهم أمام المحكمة. وسيواجه حسن عددا من ضحاياه في قاعة المحكمة لأنه سيتولى الدفاع عن نفسه. وهو يستخدم كرسيا متحركا لأنه أصيب بالشلل، عندما أطلق عليه شرطي في القاعدة العسكرية النار. Nidal Hassan, who is defending himself, admitted to killing the soldiers, claiming it was to protect Muslims and Taliban members in Afghanistan, but the military judge rejected his argument of "protecting others." If Hassan, 42 years old, is convicted of killing 13 people and injuring others, he will face the death penalty. It is considered the deadliest incident among the non-combat attacks that occurred at a U.S. military base. Witnesses stated that on November 5, 2009, he entered a clinic bustling with soldiers who were waiting for their turns for medical examinations or vaccinations, then climbed onto a desk and opened fire with two guns in his hands, only stopping to reload. Topics that may be of interest as a conclusion will be presented by the representative and the prosecution, providing evidence that Hassan leaned towards extremist ideas and was visiting sites in search of jihadists "and the Taliban, hours before the attack. Captain Hassan was set to join the American forces in Afghanistan before carrying out his attack. "Workplace violence" The U.S. Department of Defense classified the incident as "workplace violence" instead of labeling it as "terrorism," which angered the victims' families, according to BBC correspondent Nick Bryant in Fort Hood. Many of the accident's victims are expected to testify in court. Hassan will face several of his victims in the courtroom because he will be defending himself. He uses a wheelchair because he was paralyzed when a police officer shot him at the military base.
AraT5	قالت المحكمة الأمريكية إن مسلحا قتل 13 جنديا أمريكيا وجرح آخرين في هجوم على قاعدة عسكرية في أفغانستان. The American court stated that an armed assailant killed 13 American soldiers and injured others in an attack on a military base in Afghanistan.
BERT2BERT	"قالت وزارة الدفاع الأمريكية إنها سيواجه في هجوم شنه مسلحو حركة طالبان، وذلك بعد إدانته بأنه لم يكن إلى مقتل خمسة أشخاص على الأقل. من بينهم شخصا للجنود الباكستانيين الذين قتلوا في أفغانستان. عام 2011، حسب ما ذكره وكالة أنباء الدفاع عن شهود عيان. مليار دولار. بي بي سي إن قوات الشرطة لا يمكن أن يكون هناك. The U.S. Department of Defense stated that it will confront an attack launched by Taliban militants, following its condemnation of the incident that resulted in the deaths of at least five people. Among them was a person among the Pakistani soldiers who were killed in Afghanistan. In 2011, according to what was reported by the eyewitness news agency. One billion dollars. The BBC reports that the police forces cannot be present.
mBERT2mBERT	قالت الشرطة الأمريكية إنها قتلوا في هجوم بأنه لم يكن على الأقل، وذلك بعد يومين من بينهم. ا للرئيس الأمريكي دونالد ترامب إلى مقتل 30 شخصا. "ات مع حركة طالبان. تا عليه التي يسيطر عليها العسكرية الإسلامية. ون الذين يمكن أن تكون ما ذكره مسؤولون / كانون الماضي. ان كان قد أعلنه لا يقل عن. The American police stated that they killed in an attack that was not at least, just two days after among them. The American President Donald Trump commented on the killing of 30 people. "With the Taliban movement." Until it is controlled by the Islamic military. And those who could be what officials mentioned / last January. If he announced it, it is no less than.

Figure 10: A sample text that our models have summarized.

more accurate and contextually aware summarization models. By addressing these directions, future research can improve the efficiency, coherence, and applicability of summarization models across a broader range of real-world use cases.

References

[1] C. Slamet, A. Atmadja, D. Maylawati, R. Lestari, W. Darmalaksana, and M. A. Ramdhani, "Automated text summarization for indonesian article using vector space model," in *IOP Conference Series: Materials Science and Engineering*, vol. 288, p. 012037, IOP Publishing, 2018.

[2] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, and K. Kochut, "Text summarization techniques: a brief survey," *arXiv preprint arXiv:1707.02268*, 2017.

[3] L. H. Belguith, M. Ellouze, M. H. Maaloul, M. Jaoua, F. K. Jaoua, and P. Blache, "Automatic summarization," *Natural language processing of semitic languages*, pp. 371–408, 2014.

[4] A. Khan and N. Salim, "A review on abstractive summarization methods," *Journal of Theoretical and Applied Information Technology*, vol. 59, no. 1, pp. 64–72, 2014.

[5] H. N. Fejer and N. Omar, "Automatic arabic text summarization using clustering and keyphrase extraction," in *Proceedings of the 6th International Conference on Information Technology and Multimedia*, pp. 293–298, IEEE, 2014.

[6] A. Elsaid, A. Mohammed, L. F. Ibrahim, and M. M. Sakre, "A comprehensive review of arabic text summarization," *IEEE Access*, vol. 10, pp. 38012–38030, 2022.

[7] S. P. Singh, A. Kumar, H. Darbari, L. Singh, A. Rastogi, and S. Jain, "Machine translation using deep learning: An overview," in *2017 international conference on computer, communications and electronics (comptelix)*, pp. 162–167, IEEE, 2017.

[8] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1253, 2018.

[9] H. Liu, S.-G. Li, H.-X. Wang, and G.-J. Li, "Adaptive fuzzy synchronization for a class of fractional-order neural networks," *Chinese Physics B*, vol. 26, no. 3, p. 030504, 2017.

[10] M. Al-Maleh and S. Desouki, "Arabic text summarization using deep learning approach," *Journal of Big Data*, vol. 7, no. 1, p. 109, 2020.

[11] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in neural information processing systems*, vol. 27, 2014.

[12] D. Suleiman and A. Awajan, "Deep learning based abstractive text summarization: approaches, datasets, evaluation measures, and challenges," *Mathematical problems in engineering*, vol. 2020, pp. 1–29, 2020.

[13] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[14] V. Ashish, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, p. I, 2017.

[15] S. Ruder, *Neural transfer learning for natural language processing*. PhD thesis, NUI Galway, 2019.

[16] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, "Pre-trained models for natural language processing: A survey," *Science China Technological Sciences*, vol. 63, no. 10, pp. 1872–1897, 2020.

[17] W. Antoun, F. Baly, and H. Hajj, "Aragpt2: Pre-trained transformer for arabic language generation," *arXiv preprint arXiv:2012.15520*, 2020.

[18] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[19] M. Bani-Almarjeh and M.-B. Kurdy, "Arabic abstractive text summarization using rnn-based and transformer-based architectures," *Information Processing & Management*, vol. 60, no. 2, p. 103227, 2023.

[20] M. Kahla, Z. G. Yang, and A. Novák, "Cross-lingual fine-tuning for abstractive arabic text summarization," in *Proceedings of the international conference on recent advances in natural language processing (ranlp 2021)*, pp. 655–663, 2021.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[22] Y. Tang, C. Tran, X. Li, P.-J. Chen, N. Goyal, V. Chaudhary, J. Gu, and A. Fan, "Multilingual translation with extensible multilingual pretraining and finetuning," *arXiv preprint arXiv:2008.00401*, 2020.

[23] W. Antoun, F. Baly, and H. Hajj, "Arabert: Transformer-based model for arabic language understanding," *arXiv preprint arXiv:2003.00104*, 2020.

[24] E. M. B. Nagoudi, A. Elmadany, and M. Abdul-Mageed, "Arat5: Text-to-text transformers for arabic language generation," *arXiv preprint arXiv:2109.12068*, 2021.

[25] A. ELSAID, A. MOHAMMED, L. Fattouh, and M. SAKRE, "Hybrid arabic text summarization approach based on seq-to-seq and transformer," 2023.

[26] Y. M. Wazery, M. E. Saleh, A. Alharbi, A. A. Ali, *et al.*, "Abstractive arabic text summarization based on deep learning," *Computational Intelligence and Neuroscience*, vol. 2022, 2022.

[27] Y. A. AL-Khassawneh and E. S. Hanandeh, "Extractive arabic text summarization-graph-based approach," *Electronics*, vol. 12, no. 2, p. 437, 2023.

[28] D. Suleiman and A. Awajan, "Multilayer encoder and single-layer decoder for abstractive arabic text summarization," *Knowledge-Based Systems*, vol. 237, p. 107791, 2022.

[29] M. K. Eddine, N. Tomeh, N. Habash, J. L. Roux, and M. Vazirgiannis, "Arabart: a pretrained arabic sequence-to-sequence model for abstractive summarization," *arXiv preprint arXiv:2203.10945*, 2022.

[30] A. Qaroush, I. A. Farha, W. Ghanem, M. Washaha, and E. Maali, "An efficient single document arabic text summarization using a combination of statistical and semantic features," *Journal of King Saud University-Computer and Information Sciences*, vol. 33, no. 6, pp. 677–692, 2021.

[31] W. Etaiwi and A. Awajan, "Semg-ts: Abstractive arabic text summarization using semantic graph embedding," *Mathematics*, vol. 10, no. 18, p. 3225, 2022.

[32] T. Hasan, A. Bhattacharjee, M. S. Islam, K. Samin, Y.-F. Li, Y.-B. Kang, M. S. Rahman, and R. Shahriyar, "Xl-sum: Large-scale multilingual abstractive summarization for 44 languages," *arXiv preprint arXiv:2106.13822*, 2021.

[33] A. M. Zaki, M. I. Khalil, and H. M. Abbas, "Deep architectures for abstractive text summarization in multiple languages," in *2019 14th International Conference on Computer Engineering and Systems (ICCES)*, pp. 22–27, IEEE, 2019.

[34] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "mt5: A massively multilingual pre-trained text-to-text transformer," *arXiv preprint arXiv:2010.11934*, 2020.

[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[36] A. K. Vijayakumar, M. Cogswell, R. R. Selvaraju, Q. Sun, S. Lee, D. Crandall, and D. Batra, "Diverse beam search: Decoding diverse solutions from neural sequence models," *arXiv preprint arXiv:1610.02424*, 2016.

[37] S. Rothe, S. Narayan, and A. Severyn, "Leveraging pre-trained checkpoints for sequence generation tasks," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 264–280, 2020.

[38] G. Inoue, B. Alhafni, N. Baimukan, H. Bouamor, and N. Habash, "The interplay of variant, size, and task type in arabic pre-trained language models," *arXiv preprint arXiv:2103.06678*, 2021.

[39] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Multilingual denoising pre-training for neural machine translation," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 726–742, 2020.

[40] M. K. Eddine, A. J.-P. Tixier, and M. Vazirgiannis, "Barthez: a skilled pretrained french sequence-to-sequence model," *arXiv preprint arXiv:2010.12321*, 2020.

Authors

Karim Morsi graduated from the Faculty of Computer and Information Science (FCIS) at Ain Shams University. Currently, I am a Master's student specializing in Natural Language Processing (NLP) at FCIS. Additionally, I work as a Teaching Assistant at FCIS. My primary area of interest is NLP, where I focus on advancing research and applications in this field.

Fatma Najib is an Assistant Professor in the Information Systems Department at the Faculty of Computer and Information Sciences, Ain Shams University. She specializes in data streams, data mining, and real-time processing, with a focus on continuous query optimization and cloud computing. Her research aims to develop innovative solutions for efficient data management and processing in dynamic and scalable environments. As an expert in these fields.

Wedad Hussein is an Assistant Professor at the Faculty of Computer and Information Sciences, Ain Shams University. Her research interests include data mining, social network analysis, and web mining. She explores innovative techniques to analyze large-scale data from social media platforms and the web, aiming to uncover valuable insights from complex datasets. With expertise in these areas, she contributes significantly to advancing research in data mining and its applications to social networks and the web.

Rasha Ismail is a Professor at the Faculty of Computer and Information Sciences, Ain Shams University. Her research covers a broad range of topics, including data science, artificial intelligence, big data analytics, and information retrieval. She is committed to advancing the application of these technologies to address complex, real-world challenges. With her extensive expertise, she plays a crucial role in shaping the direction of research and innovation within these fields at the university.

Journal Submission

The International Journal of Computers and Their Applications is published four times a year with the purpose of providing a forum for state-of-the-art developments and research in the theory and design of computers, as well as current innovative activities in the applications of computers. In contrast to other journals, this journal focuses on emerging computer technologies with emphasis on the applicability to real world problems. Current areas of particular interest include, but are not limited to: architecture, networks, intelligent systems, parallel and distributed computing, software and information engineering, and computer applications (e.g., engineering, medicine, business, education, etc.). All papers are subject to peer review before selection.

A. Procedure for Submission of a Technical Paper for Consideration

1. Email your manuscript to the Editor-in-Chief, Dr. Ajay Bandi. Email: ajay@nwmissouri.edu.
2. Illustrations should be high quality (originals unnecessary).
3. Enclose a separate page (or include in the email message) the preferred author and address for correspondence. Also, please include email, telephone, and fax information should further contact be needed.
4. **Note**: Papers shorter than 10 pages long will be returned.

B. Manuscript Style:

1. **WORD DOCUMENT**: The text should be **double-spaced** (12 point or larger), **single column** and **single-sided** on 8.5 X 11 inch pages. Or it can be single spaced double column.

   LaTex DOCUMENT: The text is to be a double column (10 point font) in pdf format.
2. An informative abstract of 100-250 words should be provided.
3. At least 5 keywords following the abstract describing the paper topics.
4. References (alphabetized by first author) should appear at the end of the paper, as follows: author(s), first initials followed by last name, title in quotation marks, periodical, volume, inclusive page numbers, month, and year.
5. The figures are to be integrated in the text after referenced in the text.

C. Submission of Accepted Manuscripts

1. The final complete paper (with abstract, figures, tables, and keywords) satisfying Section B above in **MS Word format** should be submitted to the Editor-in-Chief. If one wished to use LaTex, please see the corresponding LaTex template.
2. The submission may be on a CD/DVD or as an email attachment(s). **The following electronic files should be included:**
   - Paper text (required).
   - Bios (required for each author).
   - Author Photos are to be integrated into the text.
   - Figures, Tables, and Illustrations. These should be integrated into the paper text file.
3. Reminder: The authors photos and short bios should be integrated into the text at the end of the paper. All figures, tables, and illustrations should be integrated into the text after being mentioned in the text.
4. The final paper should be submitted in (a) pdf AND (b) either Word or LaTex. For those authors using LaTex, please follow the guidelines and template.
5. Authors are asked to sign an ISCA copyright form (http://www.isca-hq.org/j-copyright.htm), indicating that they are transferring the copyright to ISCA or declaring the work to be government-sponsored work in the public domain. Also, letters of permission for inclusion of non-original materials are required.

Publication Charges

After a manuscript has been accepted for publication, the contact author will be invoiced a publication charge of **$500.00 USD** to cover part of the cost of publication. For ISCA members, publication charges are **$400.00 USD** publication charges are required.

Revised 2020

ISCA INTERNATIONAL JOURNAL OF COMPUTERS AND THEIR APPLICATIONS Vol. 32 No. 1, March 2024