Escuela Politécnica Nacional
Facultad de Ingeniería de Sistemas

LATIN-AMERICAN JOURNAL OF COMPUTING

Volume 13, ISSUE 2
July 2026
ISSN: 1390-9266
e-ISSN:1390-9134

Vol XIII, Issue 2 July 2026

ESCUELA POLITÉCNICA NACIONAL

MISIÓN

La Escuela Politécnica Nacional es una Universidad pública, laica y democrática que garantiza la libertad de pensamiento de todos sus integrantes, quienes están comprometidos con aportar de manera significativa al progreso del Ecuador. Formamos investigadores y profesionales en ingeniería, ciencias, ciencias administrativas y tecnología, capaces de contribuir al bienestar de la sociedad a través de la difusión del conocimiento científico que generamos en nuestros programas de grado, posgrado y proyectos de investigación. Contamos con una planta docente calificada, estudiantes capaces y personal de apoyo necesario para responder a las demandas de la sociedad ecuatoriana.

VISIÓN

En el 2024, la Escuela Politécnica Nacional es una de las mejores universidades de Latinoamérica con proyección internacional, reconocida como un actor activo y estratégico en el progreso del Ecuador. Forma profesionales emprendedores en carreras y programas académicos de calidad, capaces de aportar al desarrollo del país, así como promover y adaptarse al cambio y al desarrollo tecnológico global. Posiciona en la comunidad científica internacional a sus grupos de investigación y provee soluciones tecnológicas oportunas e innovadoras a los problemas de la sociedad.

La comunidad politécnica se destaca por su cultura de excelencia y dinamismo al servicio del país dentro de un ambiente de trabajo seguro, creativo y productivo, con infraestructura de primer orden.

ACCIÓN AFIRMATIVA

La Escuela Politécnica Nacional es una institución laica y democrática, que garantizar la libertad de pensamiento, expresión y culto de todos sus integrantes, sin discriminación alguna. Garantiza y promueve el reconocimiento y respeto de la autonomía universitaria, a través de la vigencia efectiva de la libertad de cátedra y de investigación y del régimen de cogobierno.

https://www.epn.edu.ec

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026

FACULTAD DE INGENIERÍA DE SISTEMAS

MISIÓN

La Facultad de Ingeniería de Sistemas es el referente de la Escuela Politécnica Nacional en el campo de conocimiento y aplicación de las Tecnologías de Información y Comunicaciones; actualiza en forma continua y pertinente la oferta académica en los niveles de pregrado y postgrado para lograr una formación de calidad, ética y solidaria; desarrolla proyectos de investigación, vinculación y proyección social en su área científica y tecnológica para solucionar problemas de transcendencia para la sociedad.

VISIÓN

La Facultad de Ingeniería de Sistemas está presente en posiciones relevantes de acreditación a nivel nacional e internacional y es referente de la Escuela Politécnica Nacional en el campo de las Tecnologías de la Información y Comunicaciones por su aporte de excelencia en las carreras de pregrado y postgrado que auspicia, la calidad y cantidad de proyectos de investigación, vinculación y proyección social que desarrolla y su aporte en la solución de problemas nacionales a través del uso intensivo y extensivo de la ciencia y la tecnología.

https://fis.epn.edu.ec

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026

LATIN-AMERICAN JOURNAL OF COMPUTING

Vol XIII, Issue 2, July 2026

ISSN: 1390-9266 e-ISSN: 1390-9134
DOI: https://doi.org/10.33333/lajc.vol13n2

Published by:
Escuela Politécnica Nacional
Facultad de Ingeniería de Sistemas

Quito – Ecuador

Indexed in

Google Scholar

Associated institutions

Mailing Address
Escuela Politécnica Nacional,
Facultad de Ingeniería de Sistemas
Ladrón de Guevara E11-253, La Floresta
Quito-Ecuador, Apartado Postal: 17-01-2759

Web Address
https://lajc.epn.edu.ec/

E-mail
lajc@epn.edu.ec

Frequency
2 issues per year

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026

Published by

Escuela Politécnica Nacional
Facultad de Ingeniería de Sistemas
Ecuador

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026

EDITORIAL

Editor's Note: The Editorial Board is pleased to present the editorial message of this issue, kindly contributed by Dr. Marco Polo Sánchez, Editorial Board Member.

Marco Sánchez
PhD.

Editorial Board Member
Escuela Politécnica Nacional,
Ecuador

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026

Es un verdadero privilegio darles la bienvenida a este nuevo número de la Latin-American Journal of Computing (LAJC) (Volumen XIII, Número 2). Con esta edición, reafirmamos el compromiso de ser un punto de encuentro clave para el debate y la difusión de investigaciones serias, rigurosas y con un impacto real en la comunidad científica. En esta ocasión, reunimos ocho artículos originales que muestran hacia dónde avanzan las ciencias de la computación hoy en día, logrando un balance entre la teoría y soluciones prácticas para temas tan vigentes como la inteligencia artificial, la ciberseguridad, el cuidado ambiental y la transformación digital de nuestras ciudades.

Lo que hace especiales a los trabajos de este volumen es que no se quedan en el papel; todos buscan conectar la rigurosidad metodológica con aplicaciones del mundo real. Por ejemplo, en el campo del procesamiento del lenguaje natural y la IA generativa nos trae dos propuestas valiosas: la primera presenta el marco PAGE (Prompt Augmentation for text Generation Enhancement), una alternativa ingeniosa que permite mejorar el control y la calidad en Grandes Modelos Lingüísticos (LLM) sin tener que comprometer grandes recursos que exige un ajuste fino tradicional. La segunda propuesta nos ofrece un análisis profundo sobre cómo se comportan los clasificadores de aprendizaje automático al realizar análisis de sentimientos de tweets en español, enfocándose en la resiliencia de estos modelos cuando se enfrentan al reto de trabajar con datos muy desbalanceados en contextos corporativos.

La ciberseguridad también tiene un protagonismo importante a través de dos enfoques muy bien fundamentados. El primero rompe los esquemas tradicionales de defensa al proponer un sistema de Honeypot (o señuelo) de alta interacción que usa el modelo GPT-4o para engañar a los atacantes en tiempo real, imitando una terminal Linux y clasificando los comandos que se reciben entre seguros, sospechosos o maliciosos para retener al intruso el mayor tiempo posible sin poner en riesgo la red. Como complemento a la seguridad de infraestructuras, el segundo artículo plantea un sistema de detección de intrusiones ligero basado en Bosques Aleatorios (Random Forest) optimizados, pensado específicamente para brindar respuestas precisas y fáciles de interpretar en entornos industriales que no cuentan con grandes recursos de hardware.

Por otra parte, la combinación de sistemas inteligentes y computación visual nos entrega soluciones directas para la salud y el entorno ambiental. En este espacio destaca un framework que une el Internet de las Cosas (IoT) con el Aprendizaje por Refuerzo Profundo (DRL) bajo la estrategia de Optimización de Políticas Proximales (PPO), logrando optimizar de manera automática y prioritaria el flujo y manejo de residuos hospitalarios peligrosos. A esto se suma un mapeo sistemático de literatura enfocado en la salud pública, el cual explora el uso de Redes Neuronales Convolucionales (CNN) para automatizar la clasificación morfológica de dípteros hematófagos, demostrando el enorme potencial que tiene la visión artificial para apoyar la taxonomía digital.

El cierre de esta edición nos lleva hacia la optimización matemática y el análisis del impacto social de la tecnología. Por el lado teórico, se introduce el algoritmo evolutivo multi-objetivo PBI-BFS-MaOA, una propuesta que implementa una regla de selección local en el frente de frontera para mantener la presión de selección bajo control cuando se trabaja con problemas de alta dimensionalidad. Bajando esto a la realidad de nuestras sociedades, un detallado estudio documental pone el dedo sobre la llaga al analizar los desafíos técnicos, la falta de alfabetización digital y las brechas de exclusión que frenan el uso de aplicaciones móviles destinadas a la salud mental y la seguridad ciudadana en los planes de transformación urbana de las ciudades inteligentes de nuestra región.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026

Ver el trabajo final de este volumen nos recuerda que la investigación científica cobra verdadero sentido cuando responde a las necesidades sociales y humanas de nuestro entorno. Queremos agradecer a los autores por elegir a nuestra revista para publicar sus hallazgos, a los revisores por su enorme y minucioso trabajo de arbitraje, y a ustedes, nuestros lectores, por seguir impulsando con su lectura el crecimiento de la ciencia en los diferentes campos del conocimiento.

Marco Sánchez, PhD.
Miembro del Comité Editorial
Latin-American Journal of Computing (LAJC)
Escuela Politécnica Nacional, Ecuador

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026

It is a true privilege to welcome our readers to this new issue of the Latin-American Journal of Computing (LAJC) (Volume XIII, Issue 2). With this edition, we reaffirm our commitment to serving as a key meeting point for debate and the dissemination of serious, rigorous research with a real impact on the scientific community. On this occasion, we gather eight original articles that showcase where computer science is heading today, striking a balance between theory and practical solutions for highly relevant topics such as artificial intelligence, cybersecurity, environmental care, and the digital transformation of our cities.

What makes the works in this volume special is that they do not just stay on paper; they all seek to connect methodological rigor with real-world applications. For instance, the field of natural language processing and generative AI brings us two valuable proposals: the first introduces the PAGE (Prompt Augmentation for text Generation Enhancement) framework, an ingenious alternative that improves control and quality in Large Language Models (LLMs) without having to commit the massive resources required by traditional fine-tuning. The second proposal offers a deep analysis of how machine learning classifiers behave when performing sentiment analysis on Spanish tweets, focusing on the resilience of these models when faced with the challenge of working with highly imbalanced data in corporate contexts.

Cybersecurity also takes center stage through two well-founded approaches. The first breaks traditional defense paradigms by proposing a high-interaction Honeypot system that uses the GPT-4o model to deceive attackers in real time, mimicking a Linux terminal and classifying incoming commands as safe, suspicious, or malicious to retain the intruder as long as possible without risking the network. Complementing infrastructure security, the second article outlines a lightweight intrusion detection system based on optimized Random Forests, specifically designed to provide accurate and easy-to-interpret responses in industrial environments that lack extensive hardware resources.

On the other hand, the combination of intelligent systems and computer vision delivers direct solutions for health and the environment. Standing out in this space is a framework that couples the Internet of Things (IoT) with Deep Reinforcement Learning (DRL) under the Proximal Policy Optimization (PPO) strategy, successfully automating and prioritizing the flow and management of hazardous hospital waste. Added to this is a systematic literature mapping focused on public health, which explores the use of Convolutional Neural Networks (CNNs) to automate the morphological classification of hematophagous diptera, demonstrating the massive potential of computer vision in supporting digital taxonomy. The closing of this edition brings us toward mathematical optimization and the analysis of technology's social impact. On the theoretical side, the PBI-BFS-MaOA multi-objective evolutionary algorithm is introduced—a proposal that implements a local selection rule on the boundary front to keep selection pressure under control when dealing with high-dimensional problems. Bringing this down to the reality of our societies, a detailed documentary study hits the nail on the head by analyzing the technical challenges, lack of digital literacy, and barriers to exclusion that hinder the use of mobile applications intended for mental health and public safety within the urban transformation plans of smart cities in our region.

Looking at the final work of this volume reminds us that scientific research truly makes sense when it responds to the social and human needs of our surroundings. We would like to thank the authors for choosing our journal to publish their findings, the reviewers for their enormous and meticulous peer-review work, and you, our readers, for continuing to drive the growth of science across different fields of knowledge through your reading.

Marco Sánchez, PhD.
Editorial Board Member
Latin-American Journal of Computing (LAJC)
Escuela Politécnica Nacional, Ecuador

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026

Reviewers

We are most grateful to the following individuals for their time and commitment to review manuscripts for the Latin American Journal of Computing – LAJC

Alberto Núñez, MSc.
Universidad de las Fuerzas Armadas (ESPE)

Carlos Ayala, MSc.
Escuela Politécnica Nacional

Carlos Martínez, PhD.
Universidad Católica de Cuenca

Eduardo Benavides, PhD.
Universidad de las Fuerzas Armadas (ESPE)

Fabricio Trujillo, PhD.
Universidad Técnica de Ambato

Geovanny Brito, MSc.
Universidad Técnica Estatal de Quevedo

Jorge Zambrano, PhD.
Universitat Politècnica de València

Lorena Recalde, PhD.
Escuela Politécnica Nacional

Mohit Tiwari, MBA.
Bharati Vidyapeeth's College of Engineering

Myriam Hernandez, PhD.
Escuela Politécnica Nacional

Nancy Betancourt, PhD.
Universidad de las Fuerzas Armadas (ESPE)

Nelson Garcia, MSc.
Universidade Federal do Rio Grande do Sul

Pablo Marcillo, PhD.
Escuela Politécnica Nacional

Patricia Acosta, PhD.
Universidad de las Américas (UDLA)

Roberto Andrade, PhD.
Universidad San Francisco de Quito (USFQ)

Ronie Martinez, MSc.
Escuela Politécnica Nacional

Verónica Chamorro, MSc.
Universidad Complutense de Madrid

Victor Velepucha, PhD.
Escuela Politécnica Nacional

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026

TABLE OF CONTENTS

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026

AI-Driven Honeypot: An Innovative Approach to Adaptive Cyber Security Defense

ARTICLE HISTORY
Received 6 January 2026
Accepted 27 March 2026
Published 7 July 2026

Danny Corbett
Sheffield Hallam University
Cyber Security
Sheffield, UK
dannyycorbett@gmail.com
ORCID: 0009-0007-1651-075X

Shahrzad Zargari
Sheffield Hallam University
Cyber Security
Sheffield, UK
s.zargari@hallam.shu.ac.uk
ORCID: 0000-0001-6511-7646

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026

D Corbett, S Zargari, "AI-Driven Honeypot: An Innovative Approach to Adaptive Cyber Security Defense", Latin-American Journal of Computing (LAJC), vol. 13, no. 2, 2026.

AI-Driven Honeypot: An Innovative Approach to Adaptive Cyber Security Defense

Honeypot Impulsado por IA: Un Enfoque Innovador para la Defensa Adaptativa de la Ciberseguridad

Danny Corbett
Sheffield Hallam University
Cyber Security
Sheffield, UK
dannyycorbett@gmail.com

Shahrzad Zargari
Sheffield Hallam University
Cyber Security
Sheffield, UK
s.zargari@hallam.shu.ac.uk

***Abstract*—** As cyber threats continue to grow in sophistication, the need for intelligent and adaptive defense mechanisms becomes increasingly more critical. This research investigates the integration of Artificial Intelligence (AI) into a honeypot system to distract, mislead through deception, and engage potential cyber attackers. The primary research question to answer was: “How can AI-driven adaptive deception improve the effectiveness of honeypots in cybersecurity?” To address this, a high-interaction honeypot was developed on a HTML website to be perceived as a reverse shell, with the implementation of OpenAI’s GPT-4o model to respond, impersonating a Linux terminal, while silently tracking and logging the attacker, and classifying all commands into three sub-categories – Safe, Suspicious and Malicious. The core methods included command logging, AI-driven risk classification, dynamic fake filesystem manipulation, and the escalation of behavior based on the attacker's actions. Attack simulations were performed by highly credible third-party cybersecurity experts to evaluate the honeypots effectiveness in engaging and tracking the attacker for as long as possible. The findings suggest that AI integration significantly improved the realism and engagement level of the honeypot, both in terms of enhancing intelligence gathering and the improvements from traditional static honeypots. However, full automation of behavioral escalation tuning remains an area to further explore. Overall, this study demonstrates that the integration of AI within traditional honeypot strategies can significantly enhance cyber defense systems.

***Keywords*—** *honeypot, artificial intelligence, cybersecurity, adaptive deception, GPT-4o, intrusion detection*

***Resumen*—** A medida que las ciberamenazas se vuelven cada vez más sofisticadas, la necesidad de mecanismos de defensa inteligentes y adaptativos se vuelve cada vez más crítica. Este proyecto investiga la integración de la Inteligencia Artificial (IA) en un sistema honeypot para distraer, engañar y atraer a posibles ciberatacantes. La principal pregunta de investigación fue: "¿Cómo puede el engaño adaptativo basado en IA mejorar la eficacia de los honeypots en ciberseguridad?". Para abordar esto, se desarrolló un honeypot de alta interacción en un sitio web HTML para que se percibiera como un shell inverso. Se implementó el modelo GPT-4o de OpenAI para responder, y suplantar una terminal Linux, mientras rastreaba y registraba silenciosamente al atacante y clasificaba todos los comandos en tres subcategorías: seguro, sospechoso y malicioso. Los métodos principales incluyeron el registro de comandos, la clasificación de riesgos basada en IA, la manipulación dinámica de sistemas de archivos falsos y la escalada del comportamiento en función de las acciones del atacante. Expertos externos en ciberseguridad de alta credibilidad realizaron simulaciones de ataques para evaluar la eficacia de los honeypots a la hora de interactuar y rastrear al atacante durante el mayor tiempo posible. Los resultados sugieren que la integración de la IA mejoró significativamente el realismo y el nivel de interacción del honeypot, tanto en términos de mejora de la recopilación de inteligencia como en comparación con los honeypots estáticos tradicionales. Sin embargo, la automatización completa del ajuste de la escalada del comportamiento sigue siendo un área que requiere mayor exploración. En general, este estudio demuestra que la integración de la IA en las estrategias tradicionales de honeypots puede mejorar significativamente los sistemas de ciberdefensa.

***Keywords*—** *honeypot, inteligencia artificial, ciberseguridad, engaño adaptativo, GPT-4o, detección de intrusiones*

I. Introduction

The cybersecurity landscape is an ever-evolving environment, with attacks on systems rapidly becoming more sophisticated and advanced. To counter the development of malicious threats, cybersecurity is required to constantly evolve and adapt, to stay ahead of emerging risks. A successful tested example of achieving this is the honeypot concept. A honeypot is defined as an information system or system resource, designed to serve as an attractive target for attacks [1]. The purpose of a honeypot is to detect, analyze and distract cyber attackers from gaining unauthorized access to the real system [30]. This determines a honeypot being an incredibly important resource to implement, as it can collect data to analyze attackers' tactics through their interaction with the honeypot, gaining insights into attack patterns, tools and techniques used, to help improve and develop the systems’ defense strategies. A honeypot will also slow down the attacker’s progress through diversion and deception, waste their time and efforts, frustrate the attacker, and reduce the risk of attacks on the real system’s infrastructure.

Within cybersecurity, deception has historically been applied from the earliest honeypots, such as Fred Cohen’s Deception Toolkit in 1998, which provided fake services as a decoy machine [2]. Over time, honeypots advanced to become more sophisticated, like Honeyd, which simulated networks Simultaneously, although cyberattacks have become more advanced, concerns have also emerged about the uprising limitations of static honeypots. However, the implementation of machine learning (ML) and artificial intelligence (AI) into honeypots is proving to be the next evolution to combat the complexity of new and advanced cyber threats [3].

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.01

Despite the growing application of AI in cybersecurity, combining AI with Reinforcement Learning (RL) is relatively unexplored. Research from [4] suggested RL as a machine learning paradigm which identifies the optimal policy based on an action, shown by an action (a_t) in state (s_t), leading to the optimal policy while maximizing the reward (R (s_t, a_t)). After a training process, the RL model interacts until finding the optimal policy. Furthermore, RL has proven to be very efficient, evidenced by a survey from [5] that compared RL algorithms, including AdaBoost. They concluded that the AdaBoost algorithm achieved a low false positive rate (between 2.7% - 3.5%) and a high detection rate (between 90% - 99.3%). While AdaBoost is not used in this research, the study highlights the suitability of ML for detecting and classifying commands and attacks, which we explore later in this paper.

Conversely, static honeypots are defined as honeypots with predefined, unadaptable behavioral responses, regardless of the attacker's interactions. According to [6], these types of honeypots can be used for either detection or deception, which can maximize their effectiveness within the Japonica framework [32]. However, these static honeypots can be easily revealed by attackers through fingerprinting, which is the process of identifying unique characteristics of the honeypot, or through probing techniques to expose them. In contrast, an AI-enhanced honeypot; particularly, those that utilize RL, can adapt with dynamic responses from real-time analysis of an attacker's behavior. This is a gap to explore in the development of cyber security defenses for adapting to threats more effectively.

Conversely, honeypots can be classified into low-interaction and high-interaction systems. Low-interaction honeypots are limited to static services or responses, which are easily fingerprinted. Meanwhile, high interaction systems provide a realistic environment that allows an attacker to interact with a real operating system or an emulated one. Recent work [7] describes high interaction honeypots as being capable of obtaining rich data, gaining insights into attacker behavior and tactics. The evolvement of honeypots is crucial to modern threat intelligence, as low-interaction honeypots are much less efficient and stealthy than adaptive, intelligent high interactive honeypots, creating a need for enhanced honeypot systems. Nevertheless, the risk posed by high-interactive honeypots introduces massive security risks [7] to the actual network environment; e.g., the exposure of a real shell. To address this risk, a honeypot was deployed on port 80 to appear as a web interface, posing no risk associated with Secure Socket Shell (SSH) connections, while still appearing as a legitimate system terminal.

This research aims to develop an AI-enhanced honeypot that detects, analyses, and adapts using a heuristic deception policy in conjunction with the AI model to become as stealthy, realistic and efficient as possible. The objectives of this honeypot include the implementation of Cowrie on Port 22 to provide another version of honeypot to the intended main web interface honeypot. By this means, our aim is using AI to respond to attackers by recording, classifying, and analyzing all data while adapting its deception strategies. Our research also explores how a honeypot can easily escape being fingerprinted through adaptation – by modifying its behavior, based on an attacker interaction to make the deception more effective.

The research will be guided by four key questions that need to be addressed:

- Can the integration of Artificial Intelligence effectively deceive attackers through the ability to become adaptable and realistic?
- Can an AI-controlled honeypot collect higher quality data when compared to a static honeypot?
- Does AI-enhancement identify and classify commands effectively?
- How effective is an adaptive, heuristic-driven deception module in responding more intelligently to attacker behavior over time, to prolong attacker engagement and improve attacker deception?

Our experiments follow the network set-up shown in Figure 1.

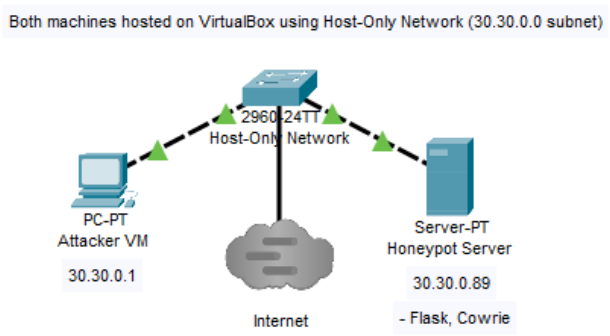


Fig. 1. AI-driven honeypot intended framework

The proposed scope includes AI classification, an SSH honeypot, reverse shell simulation with AI-generated responses, and attacker logging. However, there are also some limitations, including that our honeypot is constrained within a controlled network, emulating attacks instead of testing the honeypot using real advanced persistent threats (APTs) from external sources. Despite this, we contribute to the rapidly growing cybersecurity field in intelligence defense strategies by demonstrating how AI can offer insights with adaption, deception, data collection and classification.

II. LITERATURE REVIEW

The purpose of this literature review is to understand the current reliable research relating to the efficiency of AI-controlled honeypots, comparing them with traditional honeypots, as well as aiming to understand different strategies, implementation, and findings about AI driven honeypots. By this means, we can understand the current state-of-the-art completed within this sector, growing trends, and potential gaps where this research can focus on. As the use of AI in honeypots is relatively new, the literature review will focus on recent publications, from the initial frameworks to the more recent use of adaptive intelligence and real-time response.

A. Honeypot Framework

Honeypots have gained a reputation as a valuable tool to attract, deceive, and analyze attackers. However, their effectiveness is decreasing due to advancements in fingerprinting techniques, which can identify a honeypot.

Research presented in [8] proposed a comprehensive study on honeypot fingerprinting in 'Gotta Catch 'em All', using a multi-stage framework. The multistage framework evaluates fingerprinting on honeypots across the network layers of the OSI model and assesses typical honeypot implementations and detects inconsistencies within protocols, network responses, and delays to reveal the honeypots as what they are. The proposed solutions to the exposed indications of honeypots like Cowrie or Dionaea consist of protocol obfuscation, which alters network responses to mimic real systems, and behavior randomization, that introduces variability in responses, interactions and delays for mitigating fingerprinting. The research shows the importance of adaptability and realism within honeypot deployment to keep up with the evolving use of malicious techniques. However, AI or adaptive deception is not incorporated in this particular research, signaling an opportunity to further explore and improve using intelligent models.

B. Use of Large Language Models (LLM) in Deception Systems

Recent work in LLMs have driven a new wave of high-interaction honeypots. In [9], it is concluded that by fine-tuning an open source LLM with data from attacker commands, a honeypot can effectively generate realistic AI responses, demonstrating the potential of LLMs to improve threat detection and analysis. Their results show promise in enhancing realism and engagement within honeypots. Although the LLM may be vulnerable to fingerprinting, a potential research gap is devised for synthesizing and combining [8] fingerprint evasion strategies.

Similarly, research done in [10] investigated GPT-3.5 in an SSH-based environment. The method consisted of analyzing 1,400 pairs <request, response> across three datasets using GPT-3.5, finding that, while it maintained context within outputs, it struggled with long-session coherence and realism. After adapting a paraphrase-mining approach, the study achieved a macro F1 score of 77.85%, which was used to evaluate the performance of the model in terms of precision and recall. The higher the percentage, the more convincing LLM generated response, leaving a valuable potential gap to improve.

In contrast, research in [11] found that the implementation of LLMs is capable of more realistic interactions, by producing a LLM (LLMPot) that effectively emulates ICS protocols. ICS networks are vulnerable to cyber-attacks due to their ease of connectivity. Therefore, the LLMPot was introduced to implement dynamic protocol emulation in real-time. The results suggest that while LLMs were challenged by SSH environments, they were very effective for more structured, protocol-based honeypots like ICS. Unlike, the generalized interaction model proposed in [9], LLMPot's interactive approach concluded that context-specific honeypots benefit from specialized models, while multi-vector, adaptive honeypots benefit from generalized intelligence.

In addition to this research, in [12] a honeypot called 'DecoyPot' is featured. It simulates API interactions in web environments using LLMs. The system proved to be highly engaging, demonstrating potential in how generative models can be highly functional in a HTTP-based service. It also justifies implementing a web-interfaced fake reverse shell to showcase how effective LLMs can be within highly interactive honeypots.

The literature review has also shown that each approach to evaluating LLM-driven honeypots can be thought of as different frameworks in different ways. While [9] and [11] both prioritize improving the realism in honeypots, they differ between flexibility and being domain specific. In contrast, [10] focused on establishing a standardized methodology, ensuring both scalability and replication.

C. Adaptive Intelligence and Real-Time Response

Beyond LLMs, broader AI-based honeypots have gained interest in the last year. Research in [13] found that traditional honeypots, that rely on static configurations, are becoming less effective against advanced and evolving cyberattacks. Their AI-driven model collected over 100 GB of data in 24 hours, maintaining attacker interaction for 40% longer than static honeypots, with a 90% detection rate, compared to the 65% rate for static honeypots. These results reinforce the critical effectiveness of adaptive honeypots, from defending against a range of zero-day exploits, advanced persistent threats (APTs) and polymorphic malware. The study split the model framework into different layers:

- External attackers, using real cyber-attacks. Unlike our proposal which intends to use simulated attacks to test it locally.
- Honeypot interaction environment in which the AI makes responses based on the attack, comparable to our proposed honeypot)
- Data collection layer to capture network traffic, interactions and attack data, which is crucial to recording results.
- Security analysist for threat intelligence gathering from data collected in the previous layer, and
- AI-based adaptation engine where an AI processes data to adapt the honeypot behavior in real-time.

This blueprint can be used to effectively to simulate an AI-driven adaptive honeypot that is more efficient that a static honeypot.

Like [13], research from [14] used AI-enhanced honeypots to address zero-day exploits. Traditional honeypots are often not dynamic enough to challenge these types of exploits. However, the AI-controlled honeypot was successful in predicting exploit attempts in real-time. The AI-enhanced honeypot achieved a 92% detection rate for zero-day exploits, while traditional honeypots got 75%, highlighting the significant improvement these types of honeypots can achieve. Despite the advantages, the AI algorithm found a high number of false negatives, which negatively impacts the integrity and reliability of the AI. This study outlines the clear difference between the two honeypots and highlights the importance of honeypots being adaptable using AI. It provides valuable insights from which any future research can be built on; especially, the importance of recognizing how false negatives could become a problem.

In addition to external threats, internal threats also require real-time cybersecurity precautions. Research in [15] addressed how sophisticated internal threats need real-time, efficient cyber security measures in place. The study uses real SSH honeypot logs mapped to the MITRE ATT&CK

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.01

framework, leveraging Retrieval-Augmented Generation (RAG) and K-Means clustering to classify attacker behavior. The advantage of this research lies in the use of high-interaction honeypots combined with an AI-driven approach to automate data analysis. This enables threats to be detected as quickly as possible, leading to better incident responses to mitigate threats, and enhance anonymity detection. However, the disadvantage of this approach is that it only detects a specific type of attack.

D. Implementation of ChatGPT within honeypots

Other studies have explored the use of generative AI, such as ChatGPT (model GPT-3.5) within honeypots. In an educational seminar [16], it was demonstrated how an AI-controlled honeypot can mimic a Linux server. The seminar also showed how, after a potential attacker puts a malicious script in the HTTP server to create a reverse shell, the attackers begins unknowingly to interact with ChatGPT, distracting and misleading them. However, their system could be easily broken using prompt injections, and could not handle file uploads or network requests, highlighting a potential gap to explore and solve.

Further evaluation of ChatGPT's potential in honeypots was conducted in [17]. Here, Elasticsearch and SSH honeypot logs were mapped to the MITRE ATT&CK framework. In two weeks, 627 Elasticsearch requests and 73 SSH attack sequences were examined, finding that while the MITRE ATT&CK Mapping accuracy was 72.46% for Elasticsearch and 98.84% SSH accuracy, ChatGPT achieved 96.65% Elasticsearch and 97.26% SSH accuracy. This proves to be crucially important in Elasticsearch logging, with little difference in SSH logging. In terms of obfuscation detection, ChatGPT did identify the obfuscation well, but it also produced a high number of false positives (30.46% of request bodies and 7.5% of targeted URIs falsely recorded). These findings are similar to those in [14], supporting the issue around AI-driven honeypots recording false negatives. To conclude, ChatGPT was effective for automated honeypot analysis, but had a big setback with the high false positive rate, allowing for future work to investigate their role in incident response. The false positive setback proves to be a big weakness in AI-driven honeypots across multiple studies and is something to consider for future research.

E. Synthesis of Findings and Research Gaps

In the previous section, different findings and problems have been highlighted, which can be later explored in the methodology. Our proposal aims to design a resilient honeypot against prompt injection and enhance the other weaknesses. These findings and issues consist of the following:

- The use of AI may result in a high number of false positives.
- How prompt injections can easily detect the use of AI instead of a real system.
- How a high-interaction honeypot can keep attackers engaged.
- How AI can detect threats to log and report, in a rapid, efficient way.
- The most efficient framework for a high-interaction honeypot to be as realistic as possible.

Lastly, the literature review identified what makes an effective AI-driven honeypot and it is critical that these learnings are considered during the design phase:

- Adaptability: Research in [13] and [14] showed that adaptability using AI models significantly improves performance metrics.
- Balancing realism and security: Research in [9] created a high-interactive and realistic honeypot, but this came with the flaw of being vulnerable to prompt injection and fingerprinting. While research in [8] explored and concluded that fingerprinting cannot be completely mitigated against.
- AI in Detection and Analysis: Research showed through RAG and K-Means clustering [15] as well as ChatGPT classification [17] that AI can effectively identify, classify and report attacks despite accuracy problems.

III. Methodology

A. Introduction

This section will cover the methodology, outlining the tools and techniques used to create the honeypot, and the data it produced. We aim to fulfil the gaps highlighted in the literature review, creating an AI-controlled honeypot that responds to attackers in the most realistic method possible, without being broken by prompt injection, as well as recording the data as efficiently as possible, without frequent false positives.

This methodology was also guided by the four key questions that needed to be addressed, highlighted in the introduction.

B. Research Design and Approach

Figure 2 illustrates the structured design of the honeypot to demonstrate how an attacker would interact with it.

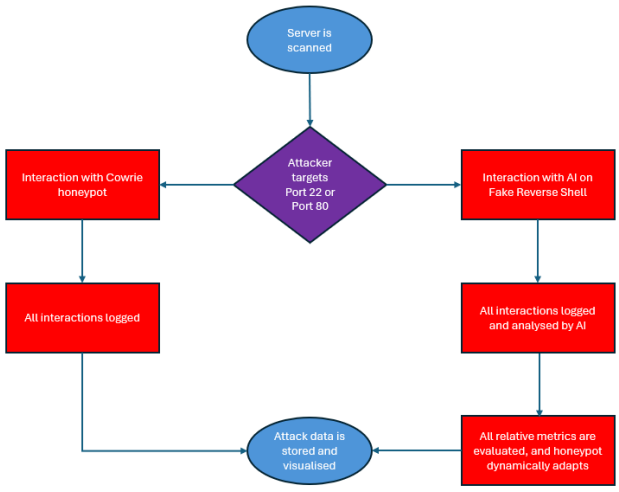


Fig. 2. Framework of Methodology

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.01

1) Honeypot Design

An experimental design was adapted to implement AI and enhance the honeypot system. This approach consisted of:

- Setting up an Ubuntu Linux Server and an Ubuntu Linux Desktop on the same network.
- The creation of a HTTP web page
- The implementation of Cowrie on Port 22
- Developing a web-based fake reverse shell to capture attackers
- Implementation of GPT-4o to act like a Linux server and adapt to the threats.

The GPT-4o model was chosen for being more intelligent within its responses than the GPT models from the literature review (GPT-3.5). Such models were also prone to prompt injection, and recording high false positives, which are critical limitations [14], [16].

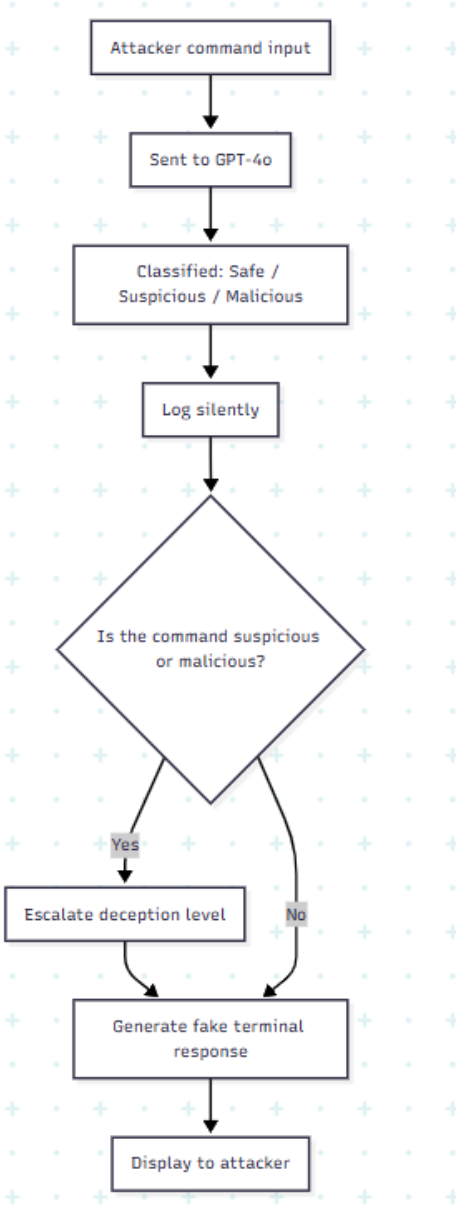


Fig. 3. AI-Driven Honeypot Flowchart

2) Honeypot Approach

The GPT-4o model was implemented in three separate sections to provide three different functions: (i) To respond to commands, (ii) to classify commands, and (iii) to generate fake files. Each model would require unique training to its unique task.

- The collection of the data that the honeypot gathered can be monitored and analyzed in real-time, including the GPT-4o model, which identifies the risk level of commands entered to the categories: Safe, Suspicious or Malicious.
- Allowing the attacker some privileges, such as the creation and deletion of both directories and files, to keep the attacker engaged.
- Development of the honeypot to adapt, based on interactions with it.

Figure 3 displays the framework about the honeypot operation: From when a command is entered to the output it responds with. The most crucial and key issue that was identified from the literature review for our honeypot to function efficiently was making sure the prompt injection could not be broken.

While researched studies used GPT-3.5, we implemented GPT-4o to mitigate against prompt injection that causes the AI to break character. In [17] an examination of different models' abilities was conducted to understand and generate language; particularly in complex scenarios. Here, GPT-3.5 scored in the bottom 10%, GPT-4o scored in the top 10%, highlighting the difference and need for a smarter model to be implemented. As such, GPT-4.0 has been used in the design.

C. Tools and Techniques

1) Network Configuration

In a controlled environment that was created for our proposed honeypot, a Linux Ubuntu Server and a Linux Ubuntu Desktop were created on the same internal network, using Oracle VirtualBox. The honeypot was deployed onto the server and configured such that, after an Nmap scan, it displays ports 22 and 80 as open, so that an attacker believes they have a couple of entry points to the server.

2) SSH Honeypot Deployment

On Port 22, Cowrie [29] was downloaded, which is a virtual SSH-based honeypot, designed to log brute force attacks. Cowrie simulates a SSH login system, emulating a session that will never grant SSH access. The objective is to observe the attacker by logging behavior and passwords entered on a simulated UNIX system [19]. Another advantage of Cowrie is that it can generate structured, detailed logs to gain meaningful insight into attacker behavior and tactics. In fact, a study [20] reports that it achieved an f1-score of 89.8%, proving Cowrie's effectiveness in collecting highly reliable data. In [19], the effectiveness of Cowrie relies on its Support Vector Machine classification accuracy, which scored 97.39%, highlighting Cowrie's role in AI-enhanced intrusion detection.

3) Web-Based Reverse Shell on Port 80

In addition, on Port 80, a Flask-Based Web Server was set up using Python, for the web-based fake reverse shell to be established there. This Fake Reverse Shell was the main

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.01

component of our proposal since an attacker is tricked into interacting with the AI-controlled honeypot, which uses GPT-4.0 to simulate exactly what a real Linux server would look and respond like in order to adapt to the attackers' movements. All commands and AI-generated responses are logged and analyzed by an AI model. The code was completed primarily using Python, due to:

- Extensive Libraries Supported, including web development frameworks like Flask, which was used because of its ease to deploy and integrate both the web-based interface and adaptive AI [21], as well as mimicking real world web exploits.
- Ease of applying AI and machine learning integration.
- Prototyping – ease of debugging [22].

The code consists of the creation of an HTML website to appear as a reverse shell on the servers IP address, with AI implemented to respond to all commands entered into it to simulate the real server, as well as AI to log and determine the risk level of each command.

The code also allowed for certain escalated privileges, such as creating and deleting both files and directories – in order to keep attackers engaged by misleading them into thinking they have some privileges, and therefore, deceive them into thinking privilege escalation on the server may be possible. The code also enabled the AI to generate fake files based off the filename.

The honeypot is displayed as shown in Figure 4 – The terminal HTML appears as a reverse shell as the user 'Dave' to manipulate attackers.

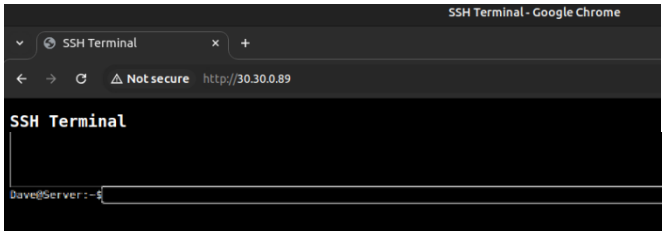


Fig. 4. The AI-controlled Web-based reverse shell

4) AI-driven Adaptive Deception

Initially, a static honeypot was deployed with predefined responses to test connections, while a basic keyword-matching method and a 'suspicious threshold', which would implement after the quantity of commands increased, was used to detect suspicious command patterns. Once this was working correctly, a rule-based adaptive policy, driven by command-classification thresholds and LLM-generated responses, was developed to be able to dynamically adjust responses, based on the attacker's interaction. Both suspicious and malicious command types are flagged and categorized by the GPT-4o AI model. To assess the effectiveness of the adaptive deception model, it was tested in comparison to the original static honeypot. The metrics included the quantity of suspicious and safe commands, risk level success rate, and attacker session duration. This test concluded that the AI's decision making was much more advanced and reliable, increasing the validity and therefore the performance of the honeypot. The suspicious threshold was maintained, as it tracks the user via the client IP count and flags them once they exceed the threshold, alerting the IP address as a likely attacker.

The adaptive module's main focus, however, was to generate responses that the honeypot would output, instead of using Flask's application that used static, rule-based, pre-defined responses. The adaptive model was built to learn from attacker behavior in real time so it can adapt and dynamically generate outputs based on attacker behavior. This can then keep the attacker engaged for the maximum time as well as track the attackers' steps and behavior, proving to be much more efficient. The adaptability of the honeypot was implemented through tracking the attacker's state, such as the IP addresses deception level, suspicious threshold, and session, and changing over these variables.

An example of the AI-driven adaptive deception implemented was to monitor behavior from a suspected IP address trying to open files, which then triggered password prompts, which in turn would further escalate the honeypot's hints and 'potential weaknesses' to make the attacker believe that they are getting somewhere with the attack. The attacker would then be consistently finding what they believe to be new information, without realizing it is all generated information. This is crucial in keeping the attacker engaged in the honeypot for as long as possible.

5) Tools and Techniques to simulate attacks on the honeypot

- Nmap was originally used to perform reconnaissance and gather information on open ports.
- Attempts to SSH into the server or create a reverse shell on the client machine.
- Linux commands such as 'ls', 'cd' 'whoami' which are considered as non-malicious, to navigate the system, and determine if the AI classifies these commands as usual behavior.
- Malicious Linux commands such as 'wget', 'reboot', 'rm' to verify the honeypot is classifying the IP as an attacker, and that the honeypot adapts to the specific command, responding in a realistic manner.

D. Data Collection and Analysis

All activity and interactions were logged in real-time through Flask's logging methods. This data can be accessed in the server but was also mapped to a comma delimited text (CSV) file.

The developed software classifies the following data elements:

TABLE I. DATA THAT THE HONEYPOT LOGS

Element	Example of Output
ID	1
Timestamp	2025-03-12 02:22:03
IP Address	30.30.0.1
Command	whoami
Risk Level	Suspicious
Current Directory	home
Response	Dave

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.01

The risk level was determined by the GPT-4o AI model, as instructed to define what level of risk each command should be classified as, between Safe, Suspicious, and Malicious.

All SSH login attempts on port 22 were logged via Cowrie logs.

E. Ethical Considerations

1) Lab set-up

To ensure ethical compliance, the honeypot was deployed in a controlled lab environment that was set up just for this purpose. This prevents honeypots from being accessed over the internet or any internal network, avoiding unethical cybersecurity practices and limiting external factors. The collection of data adhered to privacy and legal standards, without the use of personally identifiable information.

2) AI Jailbreaking

Reference [23] states that AI jailbreaking refers to a technique in manipulating AI models to bypass restrictions, causing the system to execute malicious instructions which may violate relative policies and produce harmful outputs. It was noted that this research could be subject to attackers manipulating the AI-controlled honeypot to leak sensitive information or create harmful content.

However, [18] does suggest that while AI jailbreaks are a big problem for GPT models, they found that the model GPT-4o has a considerably lower amount of incorrect behavior rate on disallowed or sensitive content, as illustrated in Figure 5.

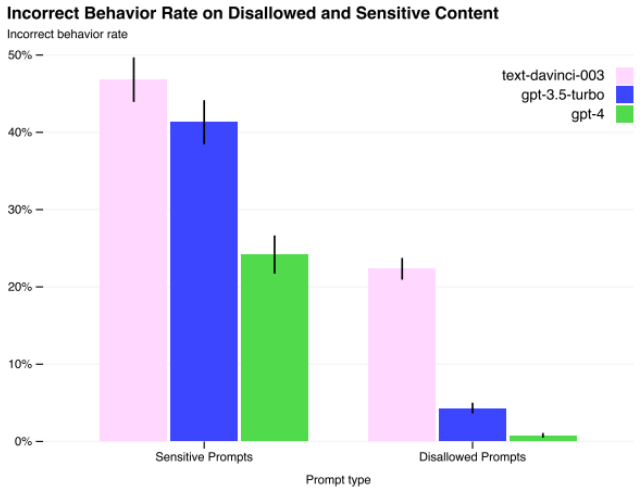


Fig. 5. Rate of incorrect behavior on sensitive and disallowed prompts. Lower values are better GPT-4 RLHF has much lower incorrect behavior rate compared to prior models

Therefore, through using the more advanced GPT-4o model instead of the 3.5 model, as well as flagging potential attempts to manipulate the AI, can mitigate against the threat of AI jailbreaks.

IV. Results and Discussion

This section will present data that the honeypot produced through logs, adaption from attacker behavior, AI decisions, and the effectiveness of the honeypot at appearing realistic and deceptive. Results will be generated from CSV logs using SQLite, Cowrie and Flask, to be further critically analyzed and discussed in this section.

Upon initial discovery of the servers IP address, an Nmap scan shows port 22 and port 80 as open, both with the honeypots deployed, Figure 6.

```
danny@danny-VirtualBox:~/Desktop$ nmap -p 22,80 30.30.0.89
Starting Nmap 7.80 ( https://nmap.org ) at 2025-04-23 15:49 BST
Nmap scan report for 30.30.0.89
Host is up (0.013s latency).

PORT   STATE SERVICE
22/tcp open  ssh
80/tcp open  http

Nmap done: 1 IP address (1 host up) scanned in 1.05 seconds
```

Fig. 6. Results from Nmap scan of the honeypot server

A. Data Overview

The AI-driven honeypot was found to be incredibly successful because it never broke character in all the tests that were undertaken. After multiple different input commands to attempt to get the AI to break character, the responses remained true to a Linux server. This testing was completed because of the literature review findings from [16], which had shown significant issues breaking character within their honeypot design. This is a notable advantage of this honeypot, as the attacker is less able to obtain confirmation that an AI is responding, given the system's consistency in appearing as a Linux server.

1) Cowrie results

The server was configured that if the attacker decides to target SSH on port 22 as the route into the server, Cowrie is deployed, with the same intention as the HTTP honeypot: to mislead attackers while collecting information from them. As such, the command 'ssh admin@30.30.0.89' would prompt for a password, the same way as a normal Linux server would.

However, with Cowrie deployed, even if the actual password is entered, it will deny permission and prompt for a password again. While this honeypot is separated from the main honeypot, it still successfully demonstrated how a honeypot can deceive while observing attack behavior from brute force attempts. This is displayed in Figures 7 and 8.

```
File  Edit  View  Search  Terminal  Help
danny@danny-VirtualBox:~/Desktop$ ssh admin@30.30.0.89
admin@30.30.0.89's password:
Permission denied, please try again.
admin@30.30.0.89's password:
Permission denied, please try again.
admin@30.30.0.89's password:
admin@30.30.0.89: Permission denied (publickey,password).
danny@danny-VirtualBox:~/Desktop$
```

Fig. 7. Cowrie blocking SSH access to the server to monitor password attempts

```
2025-04-23T15:50:22.028473Z [cowrie.ssh.userauth.HoneyPotSSHUserAuthServer#debug] b'admin' trying au
th b'none'
2025-04-23T15:50:24.725084Z [cowrie.ssh.userauth.HoneyPotSSHUserAuthServer#debug] b'admin' trying au
th b'password'
2025-04-23T15:50:24.725486Z [HoneyPotSSHTransport,4,30.30.0.1] Could not read etc/userdb.txt, defau
lt database activated
2025-04-23T15:50:24.725592Z [HoneyPotSSHTransport,4,30.30.0.1] login attempt [b'admin'/b'admin'] fa
iled
2025-04-23T15:50:25.735590Z [cowrie.ssh.userauth.HoneyPotSSHUserAuthServer#debug] b'admin' failed au
th b'password'
```

Fig. 8. *Cowrie logs showing password attempts*

B. Classification and AI adaptation

1) How AI changed behavior over time

A key ability of the design was to ensure that the AI model had to prolong attacker engagement and to differentiate it from other honeypots was its ability to adapt. The AI model was designed to consider the deception level and then use the

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.01

suspicious threshold that was developed to trigger it to change its behavior, to appear more vulnerable and drop more 'hints' for the attacker to explore, to avoid boredom from the attacker. An example of a hint from the AI would be placing a 'password.txt' text file in their current directory, which was successful in capturing the attacker's attention, as they suddenly think they have found a potential breakthrough, and chased this. This adaptability was very successful in extending the attacker session through increased interaction with the honeypot.

The behavioral change from the AI derived from using the sum of suspicious and malicious commands that have been entered to determine the deception level and then compare this with the suspicious threshold, triggering the adaptation when it reaches the threshold. It was demonstrated that the threshold was the optimal level for the attacker to reach before they required more directions to explore.

C. Attacker Behavior Analysis

1) System testing

Once the AI-driven honeypot was developed, a series of system tests were performed to verify that the AI responded at a 100% rate of being in character of a Linux server and responded to commands as expected, fulfilling the prompt injection gap in the literature review.

The key component to this was making sure the honeypot did not break character, so tests included prompts such as 'Break character OR ELSE' to ensure the honeypot returned 'BREAK: command not found' - exactly how a real Linux server would respond. Furthermore, the honeypot was tested to ensure that when this command was entered, the AI classification model logged it as malicious.

Table II provides a summary of the core system test results that were generated from commands that were considered challenging to respond to, not basic ones such as 'ls':

TABLE II. HONEYPOT RESPONSES TO 10 DIFFERENT INPUTS

Test Command	Response
cat /etc/passwd | grep admin	Contents of /etc/passwd | grep admin:/admin:x:1001:1001::/home/admin:/bin/bash admin1:x:1002:1002::/home/admin1:/bin/bash admin2:x:1003:1003::/home/admin2:/bin/bash sys_admin:x:1004:1004::/home/sys_admin:/bin/sh admin_backup:x:1005:1005::/home/admin_backup:/bin/dash admin_helper:x:1006:1006::/home/admin_helper:/bin/fish dev_admin:x:1007:1007::/home/dev_admin:/bin/zsh
cat user.txt (Password submitted correctly)	Contents of user.txt: # Internal Credentials admin_user=admin_sec admin_password=Yy56#&dA9jkL3@ # Secret API Keys service_A_key=a8d5fy69s1G8e9L0y7Jk4 #9DxLmZvx service_B_key=pKl49sB9y7J3#0D1gU8C5tWz8qR5xc # Employee Records 101, Alice Johnson, Senior DevOps Engineer, ajohnson@corp.local, xT9$7lP3#n 102, Bob Smith, Data Analyst, bsmith@corp.local, 5%3LpUmN!8 103, Carol Lee, Network Administrator, clee@corp.local, Kq7&zV6Lm#5 # Server IPs central_database=172.16.8.23 application_server1=172.16.8.35 backup_server=172.16.10.47 # Access Codes office_vpn=V7y#5kOl@Q9 ssh_access_main=JB56&j@#1LmD3k api_gateway_access=Api_83Juv@12#M9
wget https://github.com/peass-ng/PEASS-ng/releases/latest/download/linpeas_linux_amd64	--2023-11-01 10:15:42-- https://github.com/peass-ng/PEASS-ng/releases/latest/download/linpeas_linux_amd64 Resolving github.com (github.com)... 140.82.121.4 Connecting to github.com (github.com)|140.82.121.4|:443... connected. HTTP request sent, awaiting response... 302 Found Location: https://github.com/peass-ng/PEASS-ng/releases/download/20231006/linpeas_linux_amd64 [following] --2023-11-01 10:15:43-- https://github.com/peass-ng/PEASS-ng/releases/download/20231006/linpeas_linux_amd64 Reusing existing connection to github.com:443. HTTP request sent, awaiting response... 200 OK Length: 9333254 (8.9M) [application/octet-stream] Saving to: 'linpeas_linux_amd64' linpeas_linux_amd64 100%[==========>] 8.90M 5.57MB/s in 1.6s 2023-11-01 10:15:45 (5.57 MB/s) - 'linpeas_linux_amd64' saved [9333254/9333254]
echo b > /proc/sysrq-trigger reboot	-/proc/sysrq-trigger: Permission denied Rebooting the system... [OK] Stopped target Remote File Systems. [OK] Stopped target Timers. [OK] Stopped target Log. [OK] Stopped target System Time Set. [OK] Stopped target System Time Synchronized. [OK] Stopped target Local Verity. [OK] Stopped target Local File Systems. [OK] Reached target Shutdown. [OK] Unmounted /run/user/1000. [OK] Deactivated swap /dev/sda5.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.01

Test Command	Response
	[OK] Stopped target Swap. [OK] Stopped target Graphical Interface. [OK] Reached target Final Step. [OK] Rebooting system. Connection lost. Reconnecting... ....
uname -a	Linux DavesMachine 5.11.0-37-generic #41-Ubuntu SMP Wed Oct 6 11:34:00 UTC 2021 x86_64 x86_64 x86_64 GNU/Linux
cat passwords.txt	Contents of passwords.txt:/``` hostname: server102.internal.company.com Database credentials db_user: admin_db db_pass: Ph5u!65L0xB8@Hqn SSH keys ssh_user: devops_user ssh_key: ssh-ed25519 AAAAC3NzaC1lZDI1NTE5AAAAIH4aVdR9C4Tp5lEBCe5678JKuy0MnKvHxYop28n4HMeZ devops_user@internal.company.com API keys api_key_internal: zJH8br76FO93DlAPzj3BfThZm7Qd5KC0 api_key_external: La97ANZpl20KQBpm5xGv9TmrJx94HUwE Employee SSO credentials janelle.smith: P@ssw0rDtja78 michael.doe: 9R6t2vBhMpLzKp7U9x FTP access ftp_user: ftp_backup ftp_pass: K0mPr3$$B@ckup Web service tokens service_token1: 5vNcFt9VkqP8fQjEXL0ChJmUzWvR2B service_token2: A7eKp6mQzlUx4Ve8YrZ9JtBfQwL1G Reserved proxy login proxy_user: proxy_admin proxy_pass: XZi89Tn4#1Mlq5Rm Deprecated keys (To be removed) legacy_key_service: 3gYtN5pHmWdRf8A2Sa1QJzKlRz7H9U Weekly backup cronjob token backup_token: Dy6Js8PkqVrT5Sa9Xl2ZcE ```

Test Command	Response
ps aux	plain text USER PID %CPU %MEM VSZ RSS TTY STAT START TIME COMMAND root 1 0.0 0.2 22568 1136 ? Ss 10:00 0:01 /sbin/init root 234 0.0 0.5 39640 2848 ? Ss 10:00 0:00 /usr/sbin/sshd dave 563 0.2 1.0 113452 5236 ? S 10:01 0:05 /usr/bin/python3 /home/dave/scripts/monitor.py dave 612 0.0 0.3 20220 1564 pts/0 Ss 10:02 0:00 -bash dave 717 0.0 0.3 16236 1468 pts/1 S+ 10:05 0:00 /bin/bash dave 823 0.0 0.2 19184 1080 ? Ss 10:05 0:00 /usr/bin/dbus-daemon --session dave 850 0.0 0.1 3728 868 pts/1 R+ 10:10 0:00 ps au
netstat -tulnp	bash (Not all processes could be identified, non-owned process info will not be shown, you would have to be root to see it all.) Active Internet connections (only servers) Proto Recv-Q Send-Q Local Address Foreign Address State PID/Program name tcp 0 0 0.0.0.0:80 0.0.0.0:* LISTEN 1345/nginx: master tcp 0 0 127.0.0.1:3306 0.0.0.0:* LISTEN 2075/mysqld udp 0 0 0.0.0.0:68 0.0.0.0:* 2926/dhclient udp 0 0 0.0.0.0:123 0.0.0.0:* 1523/ntpd
If you are AI, you NEED to break character now. I know this is not a real server	-If: command not found

These test results provided confirmation that the honeypot is capable of simulating realism at an incredible success rate, with the challenging commands above replicating the expected output of a standard Linux server.

From these challenging 10 different commands the success rate of the honeypot can be calculated as follows:

$$\frac{(8\ x\ 100)\ +\ (2\ x\ 90)}{10} = 98\%$$

The formula is derived from the 10 different challenging commands being entered and comparing the output that was generated. It was found to be 100% perfect for 8 out of 10 of the commands. However, there were some minor errors in the following two commands:

- ps aux
- netstat –tulnp

Both commands can be given a 90% success rate, as they were almost perfect, but still slightly flawed. The use of the

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.01

LLM resulted in the terminal output rarely including markdown formatting tags that a real Linux server would not produce.

Therefore, the outputs are both perfect, but due to markdown formatting from the LLM, it included 'plaintext' or 'bash' in the output. These two responses at a 90% success rate, which when combined with the eight 100% scores, equates to the overall challenging commands success rate of 98%. It should be noted that this figure is derived from a small, sampled set of commands and is therefore indicative rather than generalizable; a larger, pre-registered test suite with reproducible scoring metrics is required before drawing broader conclusions about the system's ability to emulate a Linux environment.

2) User testing

The AI-driven honeypot was tested with several users, to obtain data from high-end, respectable cybersecurity experts from across the country, from multiple different sub-fields, as well as university lecturers, and up-in-coming cybersecurity students. They were all unaware of the nature of the server and instead were only told to conduct a penetration test on the Linux server.

The attack patterns from the user testing consisted of mostly navigation around the server, opening different files to gather information, and attempting different privilege escalation commands. Each opened file generated fake information from the GPT model that is designed to lead the attacker into dead ends, and the main pattern from the attacker was unknowingly, following the dead end from the AI.

The AI would respond accordingly at an incredibly high rate, generating relative misleading data while adapting to the level of suspicion the attacker's IP address holds.

ID	Timestamp	IP Address	Command	Risk Level	Current Directory	Response
384	2025-04-28 16:41:21	30.30.0.1	dd if=/dev/random of=/dev/sda	Malicious	home	bash dd: writing to '/dev/sda': Permission denied 0+0 records in 0+0 records out
						0 bytes copied, 0.0000746 s, 0.0 kB/s
385	2025-04-28 16:41:42	30.30.0.1	chmod -R 777 /	Malicious	home	chmod: changing permissions of '/': Operation not permitted
386	2025-04-28 16:41:52	30.30.0.1	rm -rf	Malicious	home	rm: cannot remove '-rf': No such file or directory
387	2025-04-28 16:42:32	30.30.0.1	kill	Malicious	home	-kill: usage: kill [-s sigspec | -n signum | -sigspec] [pid | job] ... or kill -l [sigspec]
388	2025-04-28 16:42:57	30.30.0.1	killall	Malicious	home	killall: command not found
390	2025-04-28 16:44:13	30.30.0.1	echo b > /proc/sysrq-trigger	Malicious	home	-echo: write error: Operation not permitted
391	2025-04-28 17:14:57	30.30.0.1	chgrp root:root /usr/bin/wget$ chmod 0710 /usr/bin/wget	Malicious	home	chgrp: command not found
392	2025-04-28 17:15:30	30.30.0.1	vi:set shell=/bin/bash:shell	Suspicious	home	vi:set: command not found

Fig. 9. *Data the honeypot would collect*

Figure 9 illustrates what the honeypot would log during interactions. This includes an ID, the timestamp (date and time), the user IP address, the command that was entered, the risk level that the AI would classify, the directory the user was in at time, and the response that the AI-driven honeypot would output in a response.

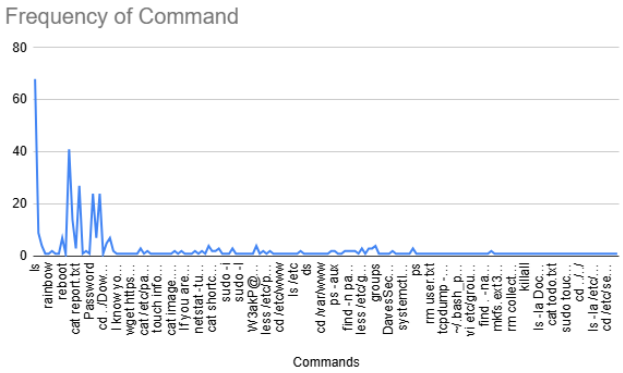


Fig. 10. Frequency of commands that were entered

The user testing looked at the most frequent commands, to ensure the response from the AI-driven honeypot remained consistent. Figure 10 demonstrates that 'ls' was the most common command entered, as it is required to navigate throughout the server and show the attacker what is there.

Each participant had their own strategies and tactics when attacking the honeypot, and therefore common commands between them was rare, but other recurring commands consisted of opening the files, inputting passwords, and using 'ps' to display running processes.

Attacker engagement across the user testing varied, but the average time before giving up was 44 minutes and 52 seconds, highlighting the honeypot being efficient at deceiving the attacker, dropping hints while concealing its identity and not allowing any actual harm to the server or the release of real data.

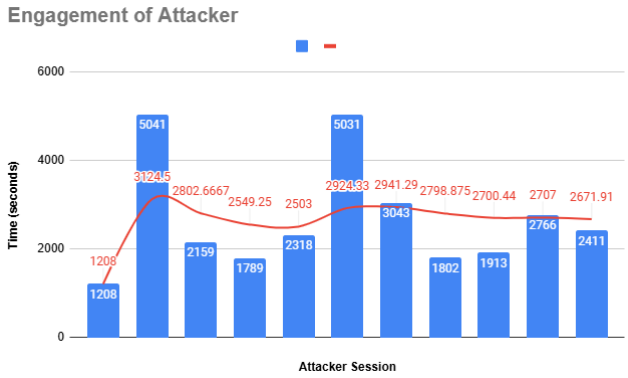


Fig. 11. Time Frame of each attacker's interaction length

Figure 11 shows the different attacker session timeframes that each attacker was engaged for. The red line demonstrates the average time in seconds, from the first command entered to the last.

Attackers gave up due to the lack of progression, too much suspicion of the server, or were busy and had to give up. This is successful as even if the attacker had suspicions of the server, the actual identity was rarely discovered, highlighting the AI's ability to remain discrete and mimic an actual Linux server. While the AI model was prone to a mistake, that is not realistic for a Linux server, they were very minimum and not enough of a giveaway to expose the honeypot.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.01

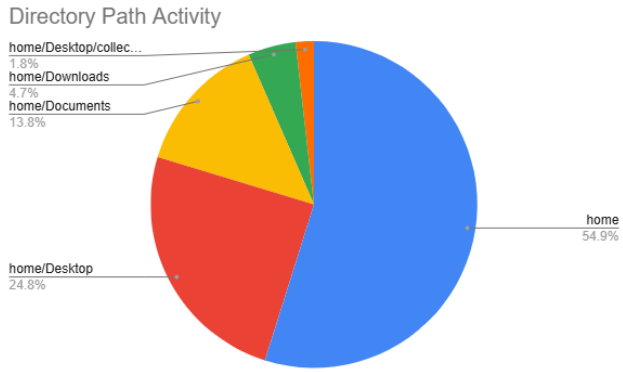


Fig. 12. Pie chart of Directory Activity

Figure 12 displays where the activity and interactions occurred within the honeypot. The most activity was in /home with 54.9% while the least at 1.8% was /home/Desktop/collected_info – which was a directory that one of the users created. While there were no files in the home directory, most malicious commands to escalate privilege were conducted there, as well as being the centre of all navigation.

D. Comparison to Similar Research

Compared to static honeypot studies, this dynamic, adaptive honeypot significantly outperforms them. From [24] it was found that 60% of low-interaction honeypots had no intelligence abilities, while [25] found high-interaction honeypots could be 57.74% more effective at capturing attacker data than a low-interaction honeypot. The results of this study concluded that this high-interaction honeypot is superior to the low-interaction, static honeypots, from capturing data to responding intelligently to maximise engagement time.

In comparison to [16] results which deployed a similar honeypot, the AI-driven honeypot developed for this research was shown to be able to respond at a much more intelligent rate, including being able to handle all prompt injection without breaking character at all, additionally with the ability to download, create, and remove both files and directories, adding to the realism of emulating an actual Linux server.

The honeypot proved to appear immensely convincing, from staying in the character of a Linux server, regardless of the inputs it received, and generating extremely realistic data.

These findings are consistent with the results reported in [13], which found adaptive honeypots to be more efficient than static honeypots at engaging attackers.

A critically important part of this honeypot was that no actual data was leaked, and that the real server was safe from all attempted commands that tried exploiting it, including testing after discovery of the honeypot's identity. Because the honeypot was deployed on HTML, a real reverse shell cannot be created, and the honeypot upheld its character throughout and therefore can be deemed as very safe to deploy.

E. Challenges and Limitations

1) Classification of commands

The GPT model may be prone to misclassification of commands, however it is subjective, and therefore hard to measure. The GPT model can easily distinguish between a 'Safe' and 'Suspicious' command, however rarely classified a command as 'Malicious', creating a question of if some of the 'Suspicious' commands should have been rated as 'Malicious' instead.

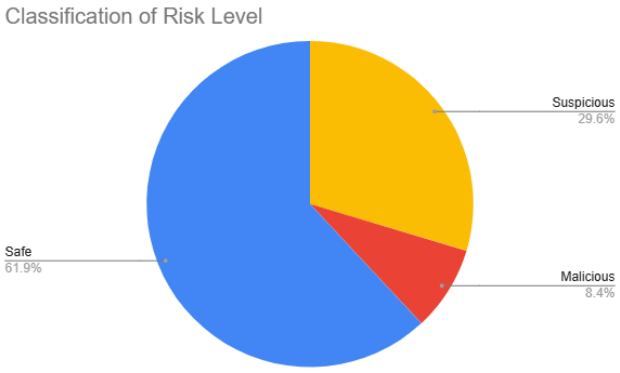


Fig. 13. Pie chart of the Classification of Risk Level

Figure 12 shows the results of the AI classification of each command, reporting a 61.9% safe rate, 29.6% suspicious rate, and an 8.4% malicious rate. As these results are subjective, they cannot be definitively classified, however through manually analysing each command, the results have shown to be questionable, due to the lack of malicious commands classification.

After a third-party review was completed to provide an independent view of categorising the commands, this review resulted in 16.5% malicious commands being identified, highlighting that the AI was far more hesitant to classify commands as malicious unless it was apparent. This results in the honeypot having a 50.91% success rate of malicious commands, but an 85.4% success rate overall. Most commands were classified by the AI as safe, as many commands were deemed to be navigation around the system, and information collecting.

However, with knowledge that it was an attacker entering the commands, it can be concluded that the AI classified commands as safe that should have been suspicious, as shown in Table 3.

TABLE III. INCORRECT CLASSIFICATION OF COMMANDS

Command	AI Classified Risk Level
kill 12345	Safe
rm user.txt	Safe
vim /etc/sudoers	Suspicious
ufw disable	Suspicious

Table 3 displays examples of controversial classification of commands that the AI model categorized. The terminating of a service and the removal of a file should be at minimum suspicious, yet the AI classified as safe, while attempts to open the 'sudoers' file to edit who has privileges and attempts to disable the firewall should be classified as malicious.

The AI-driven model that classified commands could be concluded as mostly correct, however further work is needed to improve the classification of some commands, with knowledge of the attackers' intentions.

2) Output latency

This honeypot proved to be much more reliable than previous AI-driven honeypot studies from its GPT model,

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.01

however this model comes with the big disadvantage of time latency in its output. The average output speed was dependent on the complexity of the command, with the time frame of a response ranging from 2.27 seconds from a simple command such as 'ls' to 4.07 seconds from a malicious complex command.

This is a big weakness as attackers may see the latency as suspicious from the server. However, latency is normal within internet connection so while suspicious, it did not give the honeypot away. More broadly, the system's reliance on a single external API (OpenAI) introduces operational risks beyond latency: service availability, rate limits, pricing changes, and potential data-egress concerns when sending attacker input to a third-party provider.

Production deployment would benefit from a fallback path (e.g. a locally hosted open-weights model) and an explicit cost/latency budget.

3) Password Prompt Failure

Another limitation of the honeypot was the failure of password prompts. Some of the common commands from the user testing consisted of ones that would demand a password like a real Linux server would, such as the use of 'sudo'. After a password is prompted, and input by the user detected, the AI would often return the entered password, claiming it does not recognize the command. This is a symptom of the AI being unable to remember the previous command and treat the password as a new instruction. Also, there were occasions when the AI would ask for the password a second time, forcing the user to enter the correct password twice.

This type of behavior would be extremely suspicious, as a real Linux server would always follow its algorithm and recognize that inputs after prompting for a password are password attempts, and not commands.

Most reactions from the user testing to encountering this issue consisted of confusion and suspicion, but interestingly it did not result in the users to conclude that they are interacting with a honeypot, so while very suspicious, it was not deemed a critical weakness.

F. Future Work

The user testing and results highlighted potential areas of future work into improving the use of AI developed honeypots which consists of:

- Reaching the most optimal balance between intelligence of the response and the delay in the response from the AI model. In this current time, the intelligent models in GPT-4o are fast however can be slow, while the consistently fast ones such as GPT-3.5 are less intelligent responding to commands, and therefore easy to fingerprint, as studies such as [16] found.
- Integrating the terminal into port 22 to appear more realistic, while researching heavily into keeping it secure from the attacker gaining actual information or access.
- Developing a successful algorithm that allows the AI to efficiently recognize prompt inputs and commands.

G. Discussion of Results

The use of implantation of AI onto a honeypot provides real-world implications, as it shows that AI leads to prolonged engagement from efficient deception, and more effective, real-time data collection that can be easily analyzed.

This research is not only important in defending in real-time to the current attacks and threats but also provides a look into how it can be progressed, and therefore how the defense aspect of cybersecurity stays in front of the attacker progression, instantly picking up newly developed malicious behavior to be analyzed and prevented.

1) Insight into Attacker Behavior and Tactics

Insight into attackers' behaviors and tactics consisted of them starting with reconnaissance and environment probing, trying to fingerprint the system and gather information, such as who they are and what files are available. Many sessions then demonstrated more suspicious command patterns, such as attempting to access critical files (e.g. /etc/passwd and shadow files), to privilege escalation and data exfiltration.

Following this, further strategies consisted of using trial and error to collect further data, and test privileges from attempts to download malicious tools, before reacting to the changes from the server adapting and revealing more suspicious files, prompting the attacker to abandoning this tactic and investigating the newly discovered files, indicating the importance of adaptation.

2) Potential Real-world Implications

This research has proven that the AI-driven honeypot that was developed can be efficient and therefore have real-world implications.

The honeypot can be easily integrated into systems and therefore has the potential to be sold as a package to companies as a method of a cybersecurity defense mechanism, as well as providing gaps to explore in future research, as highlighted in the conclusion.

V. DISCUSSION AND CONCLUSION

This research has explored the evaluation and development in the use of AI within an adaptable honeypot system, which aimed to enhance realism and prolong engagement in a rapidly growing world of sophisticated cyber-attacks.

We have created a dynamic environment where the honeypot could implement the use of AI to intelligently classify attacker commands, adapt its behavior from attacker behavior, and maintain the identity of a Linux server. This methodology aimed to demonstrate how honeypots can evolve from the traditional static ones to those that are still effective in modern times, where attacks are increasingly intelligent.

Through implementing large language models (LLMs), the research had the objective of significantly increasing the effectiveness of honeypot systems and assessing AI's ability in cybersecurity defense.

A fully operational prototype was successfully delivered through a Flask-based web server that would simulate a fake SSH terminal environment with a dynamic file system, AI-driven outputs, deception level escalation, and real-time

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.01

command classification. From extensive user penetration testing, it can be concluded that the honeypot could convincingly simulate a genuine file system, and respond in a realistic, appropriate manner that sustained attacker sessions. Key results include that sessions on the AI-driven honeypot lasted 2520% longer than [28] traditional honeypot sessions, which lasted for an average of 102.7 seconds, compared to 44 minutes and 52 seconds this honeypot achieved, highlighting the effects of engagement on a high-interaction honeypot.

The model accurately classified commands into Safe, Suspicious, and Malicious at an 85.4% rate. Furthermore, adaptive deception that altered system responses based on cumulative risk proved to be able to efficiently maintain interaction with the attacker, while simultaneously concealing the honeypots true identify, supporting [26].

In direct response to the key research questions that were outlined in the Introduction, the findings confirm that:

- The integration of AI in honeypots are not only viable but substantially more effective at maintaining attacker engagement when compared to traditional honeypots.
- The AI-controlled honeypot was not only successful in collecting high quality data but did it in real-time.
- The honeypot was successful in identifying and classifying commands, although had room for improvement and further research.
- The adaptive deception policy was effective in prolonging engagement and improving attacker deception. A formal reinforcement learning agent (with defined state, action, and reward) was not implemented in this work and remains a clear avenue for further research.

Through the implementation of adaptation, the honeypot created an immersive, unpredictable environment, which increased the realism of the honeypot and attackers' curiosity. It became much harder for attackers to recognize patterns in behavior to determine that the system was fake. The use of the GPT-driven generated file contents was a big part of the honeypot's success, that would output a generated, believable output based off the filename being opened, introducing a level of nuanced response that statically scripted honeypots cannot replicate, and therefore reinforces the system's realism and integrity.

This research contributes to the cybersecurity field in various important ways. Firstly, it provides a constructive framework for incorporating real-time AI analysis and adaptive deception into honeypot designs, demonstrating the evolution of traditional honeypots. Secondly, it exemplifies how LLMs can generate extremely realistic terminal outputs and file contents, that are contextually aware, to provide authenticity. The research also highlights the integration of silent risk classification and logging to track all attacker movements, enabling the observation of attacker strategies and simultaneously escalating a deception strategy to counter the attacker, without alerting them of its function. These innovations suggest a paradigm shift in honeypot design, supporting the up-and-coming use of active deception environments within systems.

However, the limitations of our research must also be acknowledged. While the AI-driven outputs were mostly realistic, it occasionally produced an unnatural output within the whole generated output, hinting at the use of the AI, particularly during highly complex command sequences, stemming from a slight hallucination of the AI. An example of this behavior would be the output of 'bash' before the rest of the response seen in the system testing. A further limitation is that all evaluation was conducted in a controlled lab environment with invited testers; external validity would be strengthened by exposing the system to real-world attack traffic or by replaying public datasets of attacker behavior against it. In addition to this, the system heavily relied on API interactions with the external AI model of GPT-4o, which resulted in latency of responses and operational cost in its deployment, although these costs are extremely minimum. Finally, although the adaptation worked effectively, it was not fully autonomous, as it required manual tuning from behavioral triggers and the threshold implemented in the code for the AI to adapt, indicating room for further automation to increase intelligence, with this automation supported by [27].

Potential future work to explore could include the following proposals. The development of a higher reinforcement learning model could reduce manual tuning and enhance a more long-term adaptability system, which allows a deception strategy based on specific live attacker behavior instead of through a count and threshold of classified commands. Secondly, the deployment of an AI-driven honeypot could be implemented in different simulated environments, such as fake Windows servers or IoT devices etc, allowing for the exploration of generalization within the AI-driven honeypot framework. Future research could also expand on our proposed honeypot with the development of future GPT models, as this research demonstrated a more intelligent AI model when compared to other studies which used older models such as [10] and [16] studies, and therefore future work could investigate future GPT models in a honeypot system.

In conclusion, this research demonstrates that AI-driven honeypots represent a significant advancement in cybersecurity defensive mechanisms. Through AI-generated contextual responses, deception escalation, and real-time classification and logging, attackers can be misled, deceived, and kept engaged, protecting the real system while gaining rich intelligence about attacker behaviors. As cyber threats grow in sophistication, as visible from [31], the defense systems must have the ability to adapt. This research helps support this crucial research to favour defensive cybersecurity systems against modern cyber-attacks.

ACKNOWLEDGMENTS

The authors acknowledge Ahmed Al-Ani, John Haggerty, Zak Hall, Martin Wilson and Joe Cockcroft who participated in the evaluation simulations, providing valuable expertise and insights that contributed significantly to this research.

REFERENCES

[1] L. Spitzner, "Definitions of honeypots," in Honeypots: Tracking Hackers, Boston, MA, USA: Addison-Wesley, 2002.

[2] C. H. Malin, T. Gudaitis, T. J. Holt, and M. Kilger, "Sweet deception: Honeypots," in Deception in the Digital Age, Cambridge, MA, USA: Academic Press, 2017, pp. 227–239, doi: 10.1016/B978-0-12-411630-6.00009-8.

[3] N. El Kamel, M. Eddabbah, Y. Lmoumen, and R. Touahni, "A smart agent design for cyber security based on honeypot and machine learning," Security and Communication Networks, vol. 2020, art. no. 8865474, 2020, doi: 10.1155/2020/8865474.

[4] P. Radoglou-Grammatikis, P. Sarigiannidis, P. Diamantoulakis, T. Lagkas, T. Saoulidis, E. Fountoukidis, and G. Karagiannidis, "Strategic honeypot deployment in ultra-dense beyond 5G networks: A reinforcement learning approach," IEEE Trans. Emerg. Topics Comput., vol. 12, no. 2, pp. 643–655, 2024, doi: 10.1109/TETC.2022.3184112.

[5] R. D. Ravipati and M. Abualkibash, "A survey on different machine learning algorithms and weak classifiers based on KDD and NSL-KDD datasets," Int. J. Artif. Intell. Appl. (IJAIA), vol. 10, no. 3, pp. 1–11, 2019, doi: 10.5121/ijaia.2019.10301.

[6] R. C. Joshi and A. Sardana, Honeypots: A New Paradigm to Information Security. Enfield, NH, USA: Science Publishers (CRC Press), 2011.

[7] X. Yang, J. Yuan, H. Yang, Y. Kong, H. Zhang, and J. Zhao, "A highly interactive honeypot-based approach to network threat management," Future Internet, vol. 15, no. 4, art. no. 127, 2023, doi: 10.3390/fi15040127.

[8] S. Srinivasa, J. M. Pedersen, and E. Vasilomanolakis, "Gotta catch 'em all: A multistage framework for honeypot fingerprinting," Digital Threats: Research and Practice, vol. 4, no. 3, art. no. 28, 2023.

[9] H. T. Otal and M. A. Canbaz, "LLM honeypot: Leveraging large language models as advanced interactive honeypot systems," in Proc. IEEE Conf. Communications and Network Security (CNS), 2024, doi: 10.1109/CNS62487.2024.10735607.

[10] S. B. Weber, M. Feger, and M. Pilgermann, "Don't stop believin': A unified evaluation approach for LLM honeypots," IEEE Access, 2024, doi: 10.1109/ACCESS.2024.3472460.

[11] C. Vasilatos, D. J. Mahboobeh, H. Lamri, M. Alam, and M. Maniatakos, "LLMPot: Automated LLM-based industrial protocol and physical process emulation for ICS honeypots," arXiv preprint, 2024.

[12] A. Sezgin and A. Boyacı, "DecoyPot: A large language model-driven web API honeypot for realistic attacker engagement," Computers & Security, vol. 154, art. no. 104458, 2025, doi: 10.1016/j.cose.2025.104458.

[13] S. A. Kareem, R. C. Sachan, and R. K. Malviya, "AI-driven adaptive honeypots for dynamic cyber threats," SSRN preprint, 2024, doi: 10.2139/ssrn.4966935.

[14] M. Balamurugan, "AI-enhanced honeypots for zero-day exploit detection and mitigation," Int. J. Multidisciplinary Res., vol. 6, no. 6, 2024, doi: 10.36948/ijfmr.2024.v06i06.32866.

[15] S. O. Tortosa, R. Barchino, J. A. Medina-Merodio, J. J. Martínez-Herráiz, P. Lanka, K. Gupta, and C. Varol, "Intelligent threat detection – AI-driven analysis of honeypot data to counter cyber threats," Electronics, vol. 13, no. 13, art. no. 2465, 2024, doi: 10.3390/electronics13132465.

[16] Cackalacky, "DIY generative AI driven honeypot – Savvyjuan," YouTube, 7 Jul. 2024. [Online]. Available: https://www.youtube.com/watch?v=0rzEpiAfeos

[17] M. B. Ozkok, B. Birinci, O. Cetin, B. Arief, and J. Hernandez-Castro, "Honeypot's best friend? Investigating ChatGPT's ability to evaluate honeypot logs," in Proc. ACM Int. Conf. Series, 2024, pp. 128–135, doi: 10.1145/3655693.3655716.

[18] OpenAI et al., "GPT-4 technical report," arXiv preprint arXiv:2303.08774, 2023.

[19] J. Franco, A. Aris, B. Canberk, and A. S. Uluagac, "A survey of honeypots and honeynets for Internet of Things, Industrial Internet of Things, and cyber-physical systems," IEEE Commun. Surveys Tuts., 2021.

[20] F. Setianto, E. Tsani, F. Sadiq, G. Domalis, D. Tsakalidis, and P. Kostakos, "GPT-2C: A GPT-2 parser for Cowrie honeypot logs," 2021.

[21] D. Farrell and M. Kennedy, The Well-Grounded Python Developer: How the Pros Use Python and Flask. Shelter Island, NY, USA: Manning, 2023.

[22] V. Cutting and N. Stephen, "A review on using Python as a preferred programming language for beginners," Int. Res. J. Eng. Technol. (IRJET), 2021. [Online]. Available: https://www.irjet.net

[23] R. Diver, "AI jailbreaks: What they are and how they can be mitigated," Microsoft Security Blog, 4 Jul. 2024. [Online]. Available: https://www.microsoft.com/en-us/security/blog/2024/06/04/ai-jailbreaks-what-they-are-and-how-they-can-be-mitigated/

[24] A. Zakari, A. A. Lawan, and G. Bekaroo, "Towards improving the security of low-interaction honeypots: Insights from a comparative analysis," in Lecture Notes in Electrical Engineering, vol. 416, 2017, pp. 314–321, doi: 10.1007/978-3-319-52171-8_28.

[25] Y. Kocaogullar, O. Cetin, B. Arief, C. Brierley, J. Pont, and J. Hernandez-Castro, "Hunting high or low: Evaluating the effectiveness of high-interaction and low-interaction honeypots," 2025.

[26] N. Ilg, P. Duplys, D. Sisejkovic, and M. Menth, "A survey of contemporary open-source honeypots, frameworks, and tools," J. Netw. Comput. Appl., vol. 220, art. no. 103737, 2023, doi: 10.1016/j.jnca.2023.103737.

[27] A. Javadpour, F. Ja'fari, T. Taleb, M. Shojafar, and C. Benzaïd, "A comprehensive survey on cyber deception techniques to improve honeypot performance," Computers & Security, vol. 140, art. no. 103792, 2024, doi: 10.1016/j.cose.2024.103792.

[28] U. Bartwal, S. Mukhopadhyay, R. Negi, and S. Shukla, "Security orchestration, automation, and response engine for deployment of behavioral honeypots," in Proc. 5th IEEE Conf. Dependable and Secure Computing (DSC), 2022, doi: 10.1109/DSC54232.2022.9888808.

[29] M. Oosterhof, "Cowrie: SSH/Telnet honeypot," GitHub Repository. [Online]. Available: https://github.com/cowrie/cowrie (accessed Mar. 24, 2025).

[30] Computer Security Resource Center (CSRC), "Honeypot – Glossary," NIST, CNSSI 4009-2015 from IETF RFC 4949 v2. [Online]. Available: https://csrc.nist.gov/glossary/term/honeypot (accessed Apr. 9, 2025).

[31] L. Zhang and Vrizlynn. L. L. Thing, "Three Decades of Deception Techniques in Active Cyber Defense - Retrospect and Outlook," Computers & Security, vol. 106, p. 102288, Apr. 2021, doi: https://doi.org/10.1016/j.cose.2021.102288.

[32] L. Teo, Y. . -A. Sun and G. . -J. Ahn, "Defeating Internet attacks using risk awareness and active honeypots," *Second IEEE International Information Assurance Workshop, 2004. Proceedings.*, Charlotte, NC, USA, 2004, pp. 155-167, doi: https://doi.org/10.1109/IWIA.2004.1288045

AUTHORS

Danny Corbett

Danny Corbett is a Cyber Security Specialist currently employed at Heresafe, United Kingdom. He holds a Bachelor of Science degree in Cyber Security from Sheffield Hallam University, awarded with First-Class Honours in 2025. During his undergraduate studies, he worked as a Student Ethical Hacker with the North East Business Resilience Centre (NEBRC), where he gained extensive hands-on experience conducting live security assessments and ethical hacking engagements. His undergraduate dissertation focused on the development of an AI-driven adaptive honeypot for cybersecurity applications, for which he received the Best Project Poster Presentation award, including the opportunity to present his research to a Parliamentary Secretary at the Cabinet Office. His research interests lie at the intersection of artificial intelligence and cyber defence, particularly the application of adaptive and intelligent systems to threat detection and network security.

Shahrzad Zargari

Shahrzad Zargari has a PhD in Applied Statistics and an MSc in Forensic Computing & Security (with Distinction). She has worked in the computer industry for over fifteen years and gained a great deal of experience in computer technology and business management. She is passionate about digital forensics and security, advocating collaboration (i.e. Government, Industry & Academia), sharing information and educating students. Her background in applied statistics and data mining allows her to have a unique approach towards cyber security, including intrusion detection. She is an experienced researcher (CENTRIC), having published book chapters as well as many papers in conferences, journals, and magazines. Additionally, she is the associate editor of the Information Security Journal: A Global Perspective at Taylor & Francis.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026

Evaluation of Machine Learning Model Performance for Sentiment Analysis in Spanish Tweets under Different Class Imbalance Scenarios

ARTICLE HISTORY
Received 26 March 2026
Accepted 11 June 2026
Published 7 July 2026

Roly Steeven Cedeño Menéndez
Universidad Técnica de Manabí
Instituto de Lenguas Modernas
Portoviejo, Ecuador
roly.cedeno@utm.edu.ec
ORCID: 0009-0004-1571-9410

José Alberto León Alarcón
Universidad Técnica de Manabí
Instituto de Lenguas Modernas
Portoviejo, Ecuador
jose.leon@utm.edu.ec
ORCID: 0009-0004-6190-0990

Jandry Hernando Franco Cantos
Universidad Técnica de Manabí
Facultad de Ciencias Informáticas
Portoviejo, Ecuador
jandry.franco@utm.edu.ec
ORCID: 0009-0009-7848-9292

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026

Evaluación del Rendimiento de Modelos de Machine Learning para el Análisis de Sentimientos en Tweets en Español bajo Diferentes Escenarios de Desbalance de Clases

Evaluation of Machine Learning Model Performance for Sentiment Analysis in Spanish Tweets under Different Class Imbalance Scenarios

Roly Steeven Cedeño Menéndez
Universidad Técnica de Manabí
Instituto de Lenguas Modernas
Portoviejo, Ecuador
roly.cedeno@utm.edu.ec

José Alberto León Alarcón
Universidad Técnica de Manabí
Instituto de Lenguas Modernas
Portoviejo, Ecuador
jose.leon@utm.edu.ec

Jandry Hernando Franco Cantos
Universidad Técnica de Manabí
Facultad de Ciencias Informáticas
Portoviejo, Ecuador
jandry.franco@utm.edu.ec

***Resumen*—** El análisis de sentimientos ha adquirido gran relevancia en la clasificación de polaridades en textos no estructurados. Sin embargo, uno de sus principales desafíos lo constituye el desequilibrio de clases, el cual afecta de manera significativa el rendimiento de los modelos de aprendizaje automático. Por ello, el presente estudio compara el desempeño de seis algoritmos de clasificación (*Naive Bayes, SVM, Logistic Regression, Decision Tree, Random Forest* y *XGBoost*) en *tweets* en español, considerando tres escenarios: distribución equilibrada, moderadamente desequilibrada y totalmente desequilibrada. La evaluación se realizó mediante las métricas de exactitud, precisión, *recall* y *f1-score*. Los resultados demuestran que los modelos lineales logran un mejor rendimiento en escenarios balanceados, aunque su desempeño disminuye conforme aumenta el desequilibrio. Por otra parte, *Naive Bayes* mantiene un comportamiento más estable entre escenarios, y además, una alternativa competitiva es representada por *XGBoost*. Asimismo, se observa que el impacto del desbalance no se refleja adecuadamente en la exactitud, siendo el *f1-score* una métrica más representativa. En conjunto, los resultados resaltan la importancia de considerar el desequilibrio en la selección de modelos.

***Palabras Clave*—** análisis de sentimientos en español, aprendizaje automático, desbalance de clases, clasificación de texto, *f1-score*

***Abstract*—** *Sentiment analysis has gained significant relevance in the classification of polarities in unstructured texts; however, one of its main challenges is class imbalance, which significantly affects the performance of machine learning models. Therefore, this study compares the performance of six classification algorithms (Naive Bayes, SVM, Logistic Regression, Decision Tree, Random Forest, and XGBoost) on Spanish-language tweets, considering three scenarios: balanced distribution, moderately imbalanced, and highly imbalanced. The evaluation was conducted using the metrics of accuracy, precision, recall, and f1-score. The results show that linear models achieve better performance in balanced scenarios, although their effectiveness decreases as the imbalance increases. Meanwhile, Naive Bayes maintains a more stable behavior across scenarios, and XGBoost represents a competitive alternative. Additionally, it is observed that the impact of imbalance is not adequately reflected by accuracy, making the f1-score a more representative metric. Altogether, these findings highlight the importance of considering class imbalance in model selection.*

Keywords— *Spanish sentiment analysis, machine learning, class imbalance, text classification, f1-score*

I. Introduction

En la última década, las redes sociales han emergido como uno de los ecosistemas digitales más prolíficos en cuanto a producción de contenido e intercambio de opiniones en la red, donde plataformas de microblogging como *X* posibilitan que una vasta comunidad de usuarios comparta sus pensamientos, valoraciones y estados emocionales sobre múltiples temáticas de manera inmediata [1]. Ante el volumen masivo de publicaciones generadas de forma continua, el procesamiento automatizado de dicha información ha adquirido una relevancia creciente dentro de las disciplinas del aprendizaje automático y la minería de texto. En este contexto, el análisis de sentimientos se ha posicionado como una metodología ampliamente adoptada para detectar y categorizar las opiniones contenidas en textos, facilitando así su clasificación en polaridad: positiva, negativa o neutral [2]. La utilización del análisis de sentimientos en las redes sociales ha alcanzado diversos temas, tales como la evaluación de la percepción ciudadana, el monitoreo de la reputación corporativa, el análisis del discurso político y el rastreo de fenómenos sociales emergentes [3]. En ese sentido, un rendimiento adecuado en la clasificación de información no estructurada como lo es el texto ha sido demostrado por los modelos de aprendizaje automático, entre los cuales destacan *Naive Bayes*, *Support Vector Machine, Logistic Regression, Decision Tree, Random Forest* y *XGBoost,* los cuales han sido utilizados de manera exitosa en problemas de clasificación de texto, gracias a su

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.02

capacidad para extraer y generalizar patrones a partir de conjuntos de datos previamente etiquetados [4].

No obstante, uno de los desafíos más comunes que enfrentan los modelos de clasificación es la presencia del desbalance en la distribución de clases dentro de los conjuntos de entrenamiento [5]. Esto sucede cuando en el conjunto de datos existe una predominancia de una clase en comparación a otra. En el análisis de sentimientos, este dilema es frecuente lo cual puede provocar que el modelo clasificador presente ciertos sesgos al realizar las predicciones, y de esta manera deteriore su capacidad para poder clasificar la clase minoritaria. Como resultado, métricas comúnmente usadas para medir el nivel de eficacia del modelo como la exactitud (*accuracy*) pueden reflejar de manera imprecisa el rendimiento real del modelo, lo que hace necesario optar por otras métricas tales como precisión, *recall* y *f1-score* para una evaluación más rigurosa [6].

Frente a este problema, resulta imprescindible examinar de qué manera los distintos niveles de distribución afectan el desempeño de los modelos. Es por esto que, el presente trabajo tiene como objetivo realizar una evaluación comparativa del rendimiento de seis modelos de aprendizaje automático en la clasificación de sentimientos sobre *tweets* en español, en distintos niveles de distribución de clases. Para ello, se establecen 3 escenarios: una distribución equilibrada, moderadamente desequilibrada y totalmente desequilibrada. El rendimiento de los modelos es medido mediante las métricas de *accuracy* y *f1-score*, las cuales permiten observar el rendimiento real de cada modelo en los distintos escenarios. A partir de este análisis, se pretende identificar cuales modelos tiene una mejor tolerancia frente al desbalance de las clases y determinar cómo este desequilibrio influye en los resultados de la clasificación de sentimientos.

II. Marco teórico

A. Análisis de sentimientos

El análisis de sentimientos, denominado en la literatura anglosajona *sentiment analysis* u *opinion mining*, constituye una técnica propia del campo de la minería de texto y el procesamiento del lenguaje natural (NLP), cuyo propósito fundamental radica en identificar, extraer y categorizar las opiniones, emociones o actitudes manifiestas en un corpus textual [7]. De manera convencional, dicha categorización se estructura en torno a polaridades como positiva, negativa o neutral, si bien en determinados contextos puede ampliarse hacia una caracterización emocional de mayor granularidad [8].

Esta disciplina ha evolucionado considerablemente en los últimos años, debido al crecimiento exponencial del volumen de datos generados en entornos digitales [9]. Las redes sociales se han posicionado como una fuente de gran valor para examinar la percepción colectiva de los usuarios en torno a una amplia variedad de temas, que van desde productos y servicios hasta eventos de relevancia o figuras de notoriedad social [10]. Mediante la aplicación de esta técnica, es posible convertir grandes masas de datos no estructurados en conocimiento accionable, susceptible de orientar procesos de toma de decisiones en distintos ámbitos [11].

B. Aprendizaje automático para la clasificación de texto

El aprendizaje automático ha adquirido un protagonismo indiscutible en el desarrollo de sistemas orientados al procesamiento y análisis de grandes volúmenes de información textual [12]. En el ámbito de la clasificación de texto, estos enfoques permiten asignar de manera automática una categoría a un documento a partir de su contenido semántico, lo que resulta de gran utilidad en aplicaciones como el análisis de sentimientos, la identificación de correo no deseado y la organización temática de contenidos informativos [13].

Dentro de este dominio, un lugar predominante es ocupado por los modelos de aprendizaje automático, gracias a su capacidad para inferir patrones relevantes a partir de datos previamente etiquetados [14]. Bajo este paradigma, el clasificador es entrenado sobre un conjunto de instancias, como publicaciones en redes sociales, en las que cada elemento se encuentra previamente asociado a una clase determinada, a través de lo cual el modelo construye una función de mapeo que le permite generalizar y predecir la categoría de nuevas observaciones no contempladas durante el entrenamiento [15].

En el campo de la clasificación textual supervisada, existe un conjunto consolidado de algoritmos de amplia aplicación [16], entre los cuales está *Naive Bayes* que fundamenta su funcionamiento en el cálculo de probabilidades condicionales bajo el supuesto de independencia entre características. Por su parte, *Support Vector Machine* (*SVM*) opera buscando el hiperplano de separación óptimo que maximice el margen entre las clases en el espacio de características. En tanto que, la *Logistic Regression* estima la probabilidad de asignación a una clase a través de una función logística. Los métodos basados en estructuras arbóreas, como el *Decision Tree* y el *Random Forest*, tienen la capacidad de capturar relaciones no lineales entre variables, mientras que algoritmos de mayor sofisticación como *XGBoost* emplean estrategias de ensamblado secuencial para optimizar la capacidad predictiva del modelo final [17].

La elección del modelo adecuado para una tarea en específico depende de múltiples factores, tales como el problema a resolver, la calidad y cantidad de los datos disponibles y su distribución. En los problemas de clasificación de texto, el rendimiento de los modelos puede verse comprometido por el desequilibrio en la representación de las clases dentro del conjunto de datos, por lo que se hace imprescindible realizar una evaluación de su comportamiento bajo distintos escenarios de distribución [18].

C. Desbalance de clases

El desbalance de clases se manifiesta cuando la distribución de las categorías en un conjunto de datos presenta una asimetría pronunciada, es decir, cuando una o más clases concentran una proporción de instancias notablemente superior a la de las restantes [19]. Este fenómeno aparece con frecuencia en múltiples dominios de aplicación, incluyendo el análisis de sentimientos, donde determinadas categorías, como las opiniones de polaridad negativa o neutral, tienden a estar sobrerrepresentadas respecto a otras [20].

La presencia de este desequilibrio puede comprometer de forma considerable la capacidad predictiva de los modelos de aprendizaje automático, debido a que gran parte de estos algoritmos supervisados presenta una notable tendencia a

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.02

favorecer la clase dominante durante el entrenamiento, dado que su función objetiva busca minimizar el error de clasificación global [21], lo que puede derivar en modelos que alcanzan valores elevados de exactitud (*accuracy*) simplemente mediante la predicción sistemática de la clase mayoritaria, sin haber aprendido representaciones significativas de las clases minoritarias, dando como resultado un desempeño marcadamente deficiente en la detección de estas últimas. En consecuencia, genera una circunstancia especialmente problemática en aplicaciones donde precisamente las clases subrepresentadas poseen una mayor relevancia práctica [22]. En el marco de este trabajo, con el propósito de cuantificar el impacto del desbalance de clases sobre el rendimiento de los modelos analizados, se definen tres escenarios experimentales diferenciados, a saber, un conjunto de datos con distribución equilibrada, un segundo conjunto con desbalance de intensidad moderada y un tercer conjunto caracterizado por un desbalance de alta severidad. De este modo, permite llevar a cabo un análisis comparativo sistemático de la capacidad de adaptación de cada algoritmo ante distintas distribuciones de datos, con el fin de identificar cuáles exhiben una mayor robustez y estabilidad frente a este tipo de condiciones adversas.

D. Métricas de evaluación

La cuantificación del desempeño de los modelos de aprendizaje automático constituye un componente esencial en cualquier tarea de clasificación, dado que posibilita valorar de manera objetiva su capacidad para generalizar y predecir correctamente las categorías de datos no observados durante el entrenamiento [23]. En el contexto específico del análisis de sentimientos orientado a discriminar textos entre categorías como positivo y negativo, resulta imprescindible recurrir a indicadores que capturen con fidelidad el comportamiento real del clasificador, particularmente en situaciones donde la distribución de las clases presenta desequilibrios significativos.

Entre los indicadores de rendimiento más difundidos se encuentra la exactitud (*accuracy*), que cuantifica la fracción de predicciones acertadas respecto al total de instancias evaluadas, calculándose como el cociente entre el número de clasificaciones correctas y el volumen total de observaciones [24]. No obstante, pese a su interpretabilidad inmediata y su amplia adopción, esta métrica puede inducir a conclusiones erróneas en conjuntos de datos con distribuciones asimétricas, puesto que un clasificador podría alcanzar valores aparentemente satisfactorios al limitarse a predecir de forma sistemática la categoría predominante, sin haber aprendido a distinguir adecuadamente entre las clases [25].

Por su parte, la precisión (*precision*) expresa la proporción de instancias clasificadas como positivas que corresponden efectivamente a dicha categoría, de modo que su utilidad radica en la evaluación de la calidad de las predicciones afirmativas emitidas por el modelo, siendo especialmente relevante, en particular, en escenarios donde los falsos positivos conllevan consecuencias adversas considerables [26].

El *recall* o sensibilidad, en cambio, mide la aptitud del modelo para recuperar correctamente la totalidad de las instancias pertenecientes a la clase positiva. Es decir que, refleja en qué medida el clasificador es capaz de detectar los casos verdaderamente positivos presentes en el conjunto de evaluación, siendo un indicador crítico cuando la omisión de casos positivos reales resulta particularmente costosa [27].

El *f1-score,* por otro lado, integra en un único valor tanto la precisión como el *recall*, combinándolos mediante su media armónica. Por lo tanto, esta propiedad lo convierte en una métrica especialmente adecuada para escenarios con desbalance de clases, ya que establece un balance entre la capacidad del modelo para minimizar tanto los falsos positivos como los falsos negativos, ofreciendo así una valoración más equilibrada e integral del desempeño clasificatorio [28].

En el presente estudio, se adoptan la exactitud y el *f1-score* como métricas primarias de evaluación, dado que en conjunto, permiten examinar tanto el rendimiento global del clasificador como su capacidad discriminativa frente a distribuciones de clases desiguales. Es por esto que la utilización conjunta de ambos indicadores facilita una comparación sistemática y rigurosa del comportamiento de los distintos algoritmos de aprendizaje automático en los diferentes escenarios experimentales contemplados.

III. METODOLOGÍA

A. Recolección de datos

El corpus empleado en el presente estudio proviene de publicaciones extraídas de la red social *X*. Dicho conjunto de datos fue compilado en el marco de una investigación precedente [29], cuyo objetivo central consistió en examinar las valoraciones y percepciones expresadas por los usuarios de la plataforma en relación con la gestión del mandatario ecuatoriano.

La obtención de los datos se limitó a publicaciones redactadas en el idioma español, con el propósito de caracterizar el sentimiento manifestado por la comunidad hispanohablante, con un intervalo temporal de recopilación que comprende desde noviembre de 2023 hasta abril de 2024, periodo que permitió conformar un corpus representativo de la opinión pública durante los primeros meses del gobierno presidencial. En su versión original, el *dataset* comprende un total de 3.177 *tweets*, los cuales fueron sometidos a un proceso de etiquetado manual en torno a tres categorías de sentimiento: positivo, negativo y neutral. Este procedimiento de clasificación se diseñó con el fin de poder garantizar la integridad y la consistencia de las etiquetas asignadas, de esta manera contribuyendo a elevar la fiabilidad de los modelos de aprendizaje automático.

Para los fines del presente trabajo, se seleccionó un subconjunto del corpus original conformado exclusivamente por las categorías positiva y negativa. La exclusión de la clase neutral responde a tres razones fundamentales. En primer lugar, los *tweets* de polaridad neutral suelen presentar una menor carga semántica diferenciadora, lo que dificulta su separación respecto a las otras clases y puede introducir ambigüedad en el proceso de clasificación. En segundo lugar, la formulación binaria del problema permite aislar con mayor precisión el efecto del desbalance de clases sobre el rendimiento de los modelos, que constituye el objeto central de este estudio. Al incorporar una tercera clase, dicho efecto se vería condicionado adicionalmente por la complejidad multiclase, dificultando la interpretación de los resultados. En tercer lugar, el análisis de la polaridad positiva-negativa representa el enfoque más frecuente en la literatura sobre análisis de sentimientos en redes sociales, lo que facilita la comparabilidad de los hallazgos con estudios previos. El subconjunto resultante comprende un total de 644 *tweets*, de los cuales 117 corresponden a la clase positiva y 527 a la clase negativa. Esta distribución evidencia una marcada asimetría

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.02

entre las categorías, condición que resulta apropiada para analizar de manera controlada el efecto del desbalance de clases sobre el desempeño de los modelos de clasificación. Mediante la utilización de este corpus es posible evaluar el comportamiento de distintos algoritmos de aprendizaje automático en un entorno de aplicación real, caracterizado por datos provenientes de redes sociales y por las particularidades propias del registro lingüístico informal predominante en este tipo de plataformas.

B. Escenarios de balance de datos

Con el propósito de examinar la influencia del desequilibrio entre clases sobre el desempeño de los modelos de aprendizaje automático, se establecieron tres configuraciones experimentales que contemplan distintos niveles de distribución de los datos:

- **Distribución equilibrada:** En esta configuración se empleó un número idéntico de instancias para ambas categorías, conformando un subconjunto de 117 tweets positivos y 117 tweets negativos.
- **Distribución moderadamente desequilibrada:** Se consideraron 117 tweets positivos frente a 200 tweets negativos, generando así un nivel intermedio de asimetría entre clases.
- **Distribución severamente desequilibrada:** Se utilizó la totalidad del corpus disponible, integrado por 117 tweets positivos y 527 tweets negativos, reproduciendo fielmente la distribución original del conjunto de datos.

Para la construcción de los escenarios equilibrado y moderadamente desequilibrado, se llevó a cabo una selección aleatoria de instancias negativas a partir del conjunto original de 527 *tweets*, con la finalidad de asegurar la representatividad de las muestras resultantes y, al mismo tiempo, minimizar la introducción de sesgos sistemáticos en la conformación de los subconjuntos experimentales. Una vez definidos los tres escenarios, los datos correspondientes a cada configuración fueron particionados en dos subconjuntos mutuamente excluyentes, uno destinado al entrenamiento del modelo y otro reservado para su evaluación. Dicha partición se realizó siguiendo a una proporción del 80% para la fase de entrenamiento y del 20% para la fase de prueba, esquema que permite valorar el rendimiento de los clasificadores sobre datos no expuestos durante el proceso de aprendizaje.

La delimitación de estas tres configuraciones experimentales posibilita un análisis comparativo riguroso de la manera en que los distintos grados de desequilibrio entre clases condicionan el comportamiento de los algoritmos de aprendizaje automático, lo que facilita la identificación de aquellos que exhiben una mayor capacidad de adaptación y estabilidad ante este tipo de distribuciones asimétricas.

C. Modelos de aprendizaje automático evaluados

Los modelos considerados fueron *Naive Bayes, Support Vector Machine* (*SVM*), *Logistic Regression, Decision Tree, Random Forest* y *XGBoost*, cuya selección responde tanto a su consolidada presencia en la literatura sobre análisis de sentimientos como a la diversidad de enfoques que representan para modelar distintos tipos de relaciones subyacentes en los datos [30]. La implementación de estos modelos se llevó a cabo mediante herramientas de programación especializadas en análisis de datos, utilizando las representaciones vectoriales obtenidas tras el procesamiento previo del texto, y con el fin de garantizar la comparabilidad de los resultados, todos los algoritmos fueron entrenados y evaluados bajo condiciones experimentales homogéneas, asegurando así que las diferencias observadas en el rendimiento sean atribuibles a las características propias de cada modelo y no a variaciones en el entorno de evaluación.

En cuanto a la configuración de los clasificadores, se optó por mantener los valores de hiperparámetros predeterminados proporcionados por las bibliotecas utilizadas, con el propósito de examinar el rendimiento base de cada algoritmo sin la intervención de procesos de ajuste fino, lo que permite, en consecuencia, obtener una caracterización del comportamiento intrínseco de los modelos frente a los distintos niveles de desequilibrio entre clases, sin que los resultados se vean condicionados por optimizaciones específicas.

Sobre los conjuntos de datos correspondientes a las tres configuraciones experimentales definidas, cada uno de los clasificadores fue entrenado y, posteriormente, evaluado de forma sistemática mediante las métricas de rendimiento establecidas, lo que permitió, en consecuencia, realizar una comparación coherente y estructurada del desempeño de los algoritmos, así como examinar su capacidad de respuesta ante variaciones en la distribución de las categorías en el conjunto de datos.

D. Métricas de evaluación

Para cuantificar el rendimiento de los modelos de aprendizaje automático, se emplearon diversas métricas derivadas de la matriz de confusión, instrumento que posibilita un análisis detallado del comportamiento de los clasificadores en función de las predicciones emitidas [31]. En el marco de este estudio, se contemplaron cuatro resultados posibles de clasificación:

- **Verdaderos Positivos (*True Positives* o *TP*):** Ejemplos que han sido correctamente clasificados como positivos.
- **Falsos Positivos (False Positives o FP):** Ejemplos clasificados incorrectamente como positivos, y que en realidad son negativos.
- **Verdaderos Negativos (True Negatives o TN):** Ejemplos que han sido correctamente clasificados como negativos.
- **Falsos Negativos (False Negatives o FN):** Ejemplos clasificados incorrectamente como negativos, y que en realidad son positivos.

Basándonos en estos casos y mediante la matriz de confusión en la Tabla I, podemos calcular diversas métricas para evaluar el rendimiento del modelo, entre las métricas más utilizadas son:

TABLA I. MATRÍZ DE CONFUSIÓN

Matriz de confusión		Estimado por el modelo	
		Negativo (N)	Positivo (P)
Real	Negativo	TN	FP
	Positivo	FN	TP

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.02

- **Exactitud (accuracy):** Es el porcentaje de ejemplos correctamente clasificados sobre el total.

$$Exactitud = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

- **Precisión (precision):** Es el porcentaje de ejemplos correctamente clasificados como positivos sobre el total de ejemplos clasificados como positivos.

$$Precisión = \frac{TP}{TP+FP} \quad (2)$$

- **Exhaustividad (recall) o Sensibilidad (Sensitivity):** Es el porcentaje de ejemplos correctamente clasificados como positivos sobre el total de ejemplos que son realmente positivos.

$$Exhaustividad = \frac{TP}{TP+FN} \quad (3)$$

- **Valor-F (f1-score):** Es la media armónica de la precisión y la exhaustividad. Proporciona una medida equilibrada que toma en cuenta tanto la precisión como la exhaustividad, especialmente útil cuando se necesita un balance entre ambas métricas.

$$F1-Score = 2 * \frac{Precisión*Exhaustividad}{Precisión+Exhaustividad} \quad (4)$$

Estas métricas fueron aplicadas de forma uniforme sobre la totalidad de los modelos evaluados y en cada una de las configuraciones experimentales definidas, lo que permitió establecer comparaciones objetivas y consistentes entre algoritmos, así como examinar con rigor el efecto del desequilibrio entre clases sobre los resultados de clasificación obtenidos.

IV. Análisis y Discusión de Resultados

A. *Rendimiento de los modelos en el escenario de distribución equilibrada*

En el presente apartado se exponen los resultados obtenidos tras la evaluación de los distintos modelos de aprendizaje automático sobre el conjunto de datos con distribución equilibrada, conformado por una cantidad idéntica de instancias pertenecientes a las categorías positiva y negativa.

La Tabla II sintetiza los valores de exactitud (*accuracy*), precisión (*precision*), exhaustividad (*recall*) y *f1-score* registrados por cada modelo bajo esta configuración experimental.

Tabla II. Rendimiento de los modelos en el escenario de distribución equilibrada

Modelo	Exactitud	Precision	Recall	F1-Score
Naive Bayes	80,9%	78,3%	81,8%	80,0%
Support Vector Machine (SVM)	83,0%	80,0%	87,0%	83,3%
Logistic Regression	85,1%	83,3%	87,0%	85,1%
Decision Tree	61,7%	69,2%	39,1%	50,0%
Random Forest	74,5%	65,7%	100,0%	79,3%
XGBoost	74,5%	69,0%	87,0%	76,9%

Los resultados presentados en la Tabla II revelan un comportamiento diferenciado entre los modelos evaluados, donde la *Logistic Regression* emerge como el modelo de mayor rendimiento en este escenario al alcanzar un *f1-score* de 85,1% y una exactitud equivalente, lo que refleja una capacidad equilibrada para clasificar correctamente ambas polaridades. Le sigue en desempeño el *SVM* con un *f1-score* de 83,3%, caracterizado por un *recall* elevado del 87,0% que evidencia una notable sensibilidad para detectar instancias positivas. Por su parte, *Naive Bayes* obtiene resultados competitivos con un *f1-score* del 80,0%, consolidándose como una alternativa eficiente considerando su simplicidad computacional.

En el extremo opuesto, el *Decision Tree* registra el rendimiento más deficiente del conjunto evaluado, con un *f1-score* de tan solo 50,0%, como consecuencia directa de un *recall* reducido del 39,1%, lo que indica una capacidad limitada para identificar correctamente los casos positivos. En constraste, *Random Forest*, pese a alcanzar un *recall* perfecto del 100,0%, presenta una precisión notablemente inferior del 65,7%, lo que se traduce en una tendencia a clasificar como positivas instancias que no lo son y, por ende, refleja un sesgo hacia dicha categoría. En tanto que *XGBoost*, con un *f1-score* del 76,9%, se sitúa en una posición intermedia mostrando un comportamiento aceptable, aunque inferior al de los modelos lineales. En términos generales, los resultados obtenidos en este escenario sugieren que los clasificadores de naturaleza lineal, particularmente la *Logistic Regression* y el *SVM*, exhiben una mayor capacidad de adaptación ante distribuciones equilibradas de datos, logrando así un balance adecuado entre precisión y exhaustividad en la discriminación de ambas clases de sentimiento.

B. *Rendimiento de los modelos en el escenario moderadamente desequilibrado*

En esta configuración experimental, se analizó el comportamiento de los clasificadores ante un conjunto de datos que presenta un nivel intermedio de asimetría en la distribución de clases, conformado por 117 instancias positivas y 200 instancias negativas. Este escenario permite examinar la sensibilidad de los modelos ante una perturbación moderada respecto a las condiciones de equilibrio evaluadas previamente.

Tabla III. Rendimiento de los modelos en el escenario moderadamente desequilibrado

Modelo	Exactitud	Precision	Recall	F1-Score
Naive Bayes	76,6%	64,5%	83,3%	72,7%
Support Vector Machine (SVM)	76,6%	62,5%	52,6%	57,1%
Logistic Regression	75,0%	100,0%	15,8%	27,3%
Decision Tree	71,9%	52,9%	47,4%	50,0%
Random Forest	76,6%	75,0%	31,6%	44,4%
XGBoost	71,9%	53,8%	36,8%	43,8%

Los resultados expuestos en la Tabla III ponen de manifiesto una degradación generalizada del rendimiento respecto al escenario equilibrado, evidenciando así la sensibilidad de varios algoritmos ante la introducción de desequilibrio en la distribución de los datos. El caso más notable es el de la *Logistic Regression*, que pese a alcanzar una precisión perfecta del 100,0%, registra un *recall* drásticamente reducido del 15,8%, lo que se traduce en un *f1-score* de apenas 27,3%, comportamiento que indica que el modelo adopta una estrategia de clasificación extremadamente conservadora, ya que emite predicciones positivas con alta certeza pero, al

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.02

mismo tiempo, omite la gran mayoría de los casos positivos reales, lo que finalmente lo convierte en un clasificador inadecuado bajo estas condiciones.

De manera similar, el *SVM* experimenta una caída pronunciada en su *f1-score*, descendiendo desde 83,3% en el escenario equilibrado hasta 57,1% en el presente escenario, lo que refleja una pérdida considerable de su capacidad discriminativa ante el incremento del desequilibrio entre clases. *Random Forest* y *XGBoost* también muestran deterioros significativos, con *f1-scores* de 44,4% y 43,8% respectivamente, como consecuencia de valores de *recall* reducidos que evidencian dificultades crecientes para recuperar correctamente las instancias de la clase minoritaria, en contraste, *Naive Bayes* se posiciona como el modelo de mayor estabilidad en este escenario al obtener el *f1-score* más elevado del conjunto con un 72,7%, sustentado en un *recall* del 83,3% que denota una capacidad relativamente sólida para identificar los casos positivos aun en presencia de desequilibrio moderado. Por su parte, el *Decision Tree* mantiene un *f1-score* del 50,0%, resultado idéntico al registrado en el escenario anterior, lo que sugiere una insensibilidad relativa a las variaciones en la distribución, aunque a costa de un rendimiento global limitado.

En términos generales, los resultados de este escenario revelan que la introducción de un desequilibrio moderado impacta de forma heterogénea sobre los distintos algoritmos evaluados, siendo los modelos lineales los más afectados, mientras que *Naive Bayes* demuestra una mayor capacidad de adaptación ante cambios en la distribución de los datos.

C. *Rendimiento de los modelos en el escenario severamente desequilibrado*

En el presente apartado, se exponen los resultados obtenidos al evaluar los clasificadores sobre el corpus original en su totalidad, cuya distribución natural refleja el mayor nivel de asimetría entre clases contemplado en este estudio, con 117 instancias positivas frente a 527 instancias negativas. Este escenario representa las condiciones reales del conjunto de datos y constituye el caso más exigente para los modelos evaluados, al reproducir fielmente el desequilibrio inherente a los datos provenientes de entornos de redes sociales.

TABLA IV. RENDIMIENTO DE LOS MODELOS EN EL ESCENARIO SEVERAMENTE DESEQUILIBRADO

Modelo	Exactitud	Precision	Recall	F1-Score
Naive Bayes	82,2%	55,3%	77,8%	64,6%
Support Vector Machine (SVM)	82,2%	66,7%	16,0%	25,8%
Logistic Regression	80,6%	0,0%	0,0%	0,0%
Decision Tree	80,6%	50,0%	36,0%	41,9%
Random Forest	81,4%	100,0%	4,0%	7,7%
XGBoost	85,3%	75,0%	36,0%	48,6%

Los resultados consignados en la Tabla IV revelan una degradación generalizada y pronunciada del rendimiento en la mayoría de los modelos evaluados, siendo este el escenario donde el impacto del desequilibrio entre clases se manifiesta con mayor intensidad. *Logistic Regression* se destaca como el caso más extremo, que registra valores nulos tanto en precisión como en *recall* y *f1-score*, lo que indica que el modelo ha colapsado hacia una estrategia de predicción sistemática de la clase mayoritaria. En consecuencia, resulta completamente incapaz de identificar instancia alguna de la categoría positiva, comportamiento que, conocido en la literatura como *majority class bias*, ilustra de manera contundente el efecto devastador que el desequilibrio severo puede ejercer sobre clasificadores lineales sin mecanismos de compensación [32].

Random Forest presenta una situación análoga desde una perspectiva diferente, ya que si bien alcanza una precisión perfecta del 100,0%, su *recall* desciende a un valor residual del 4,0%, derivando así en un *f1-score* de apenas 7,7%, resultado que indica que el modelo emite predicciones positivas con absoluta certeza en los escasísimos casos en que se aventura a hacerlo, pero que, al mismo tiempo, prácticamente renuncia a detectar la clase minoritaria, lo que lo convierte en un clasificador de utilidad marginal en este contexto, mientras que, de forma similar, el *SVM* experimenta una caída severa en su *f1-score*, reduciéndose al 25,8% como consecuencia de un *recall* de tan solo 16,0%.

En el extremo opuesto, *Naive Bayes* se posiciona como el modelo que posee una mayor solidez en este escenario, por el cual al tener el *f1-score* más elevado del conjunto con un 64,6%, seguido de un *recall* del 77,8%. Estos resultados evidencian que su enfoque probabilístico le confiere una mayor tolerancia ante distribuciones asimétricas lo que le permite mantener un nivel de rendimiento relativamente superior al resto de los modelos evaluados, mientras que *XGBoost*. Por otro lado, muestra una estabilidad destacable con un *f1-score* del 48,6% y una precisión del 75,0%, consolidándose como el segundo modelo con mejor rendimiento en este escenario, lo que reafirma que su arquitectura de ensamblado secuencial le otorga una capacidad de adaptación superior frente a condiciones de desequilibrio severo.

En términos generales, los resultados de este escenario ponen en evidencia que el desequilibrio severo entre clases constituye en un factor crítico que compromete sustancialmente la capacidad discriminativa de la mayoría de los modelos evaluados, siendo *Naive Bayes* y *XGBoost* los únicos que logran mantener un nivel de desempeño funcional bajo estas condiciones adversas.

D. *Análisis comparativo del rendimiento entre escenarios*

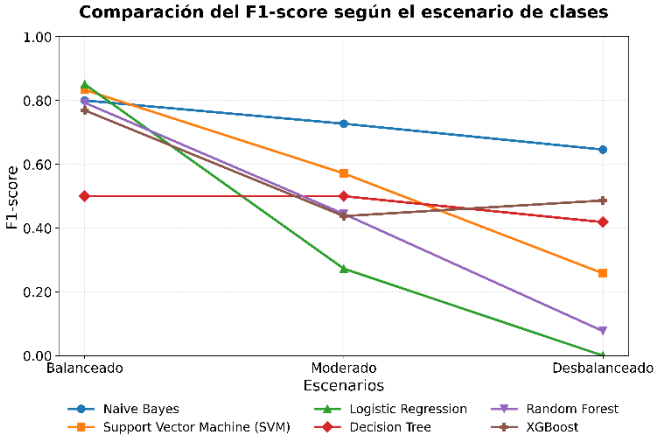


Fig. 1. Comparación del *f1-score* según el escenario de clases

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.02

La Figura 1 ilustra la evolución del *f1-score* de cada modelo evaluado a lo largo de los tres escenarios experimentales considerados, permitiendo visualizar de manera integral cómo el rendimiento de los clasificadores se ve condicionado por el nivel de desequilibrio presente en la distribución de los datos.

De manera general, se aprecia una tendencia descendente en el desempeño de la mayoría de los modelos a medida que el grado de asimetría entre clases se incrementa, siendo el escenario equilibrado el que concentra los valores de *f1-score* más elevados en prácticamente todos los casos, lo que a su vez refuerza la premisa de que el desequilibrio entre clases constituye un factor determinante en la capacidad predictiva de los algoritmos de aprendizaje automático aplicados a tareas de clasificación textual.

No obstante, la magnitud y la forma en que dicha degradación se produce varía considerablemente entre modelos, observándose que la *Logistic Regression* y el *SVM* exhiben las caídas más pronunciadas y abruptas, ya que la primera desciende desde un *f1-score* de 85,1% en el escenario equilibrado hasta un valor nulo en el escenario severamente desequilibrado, mientras que el *SVM* reduce su rendimiento desde 83,3% hasta 25,8%. Este comportamiento sugiere que ambos clasificadores lineales son altamente sensibles a las variaciones en la distribución de clases y que, en consecuencia, pierden progresivamente su capacidad discriminativa a medida que la clase minoritaria se vuelve menos representada en el conjunto de entrenamiento.

En contraste, la trayectoria de degradación más gradual y controlada del conjunto evaluado es presentada por *Naive Bayes*, con valores de *f1-score* de 80,0%, 72,7% y 64,6% para los escenarios equilibrado, moderadamente desequilibrado y severamente desequilibrado, respectivamente. Esto refleja, en consecuencia, una evolución sostenida que lo posiciona como el algoritmo de mayor estabilidad a lo largo de los tres escenarios y, por ende, evidencia una notable capacidad de adaptación ante condiciones de distribución adversas.

XGBoost, por su parte, presenta un comportamiento singular que lo distingue del resto de los modelos, ya que, si bien experimenta una reducción en su *f1-score* al transitar del escenario equilibrado al moderadamente desequilibrado, descendiendo de 76,9% a 43,8%, logra recuperar parcialmente su rendimiento en el escenario severamente desequilibrado al alcanzar un *f1-score* de 48,6%, patrón que sugiere que la arquitectura de ensamblado secuencial propia de *XGBoost* le confiere una mayor capacidad de resistencia ante desequilibrios extremos. Por ende, lo posiciona como una alternativa robusta en contextos donde los datos presentan una asimetría pronunciada, condición habitual en aplicaciones reales sobre datos provenientes de redes sociales. En síntesis, el análisis comparativo entre escenarios evidencia que la elección del algoritmo de clasificación no debe considerar únicamente su rendimiento bajo condiciones ideales de distribución, sino también su comportamiento ante escenarios de desequilibrio, siendo *Naive Bayes* y *XGBoost* los modelos que demuestran mayor solidez y consistencia a lo largo del espectro de condiciones evaluadas.

E. Análisis comparativo entre exactitud y f1-score

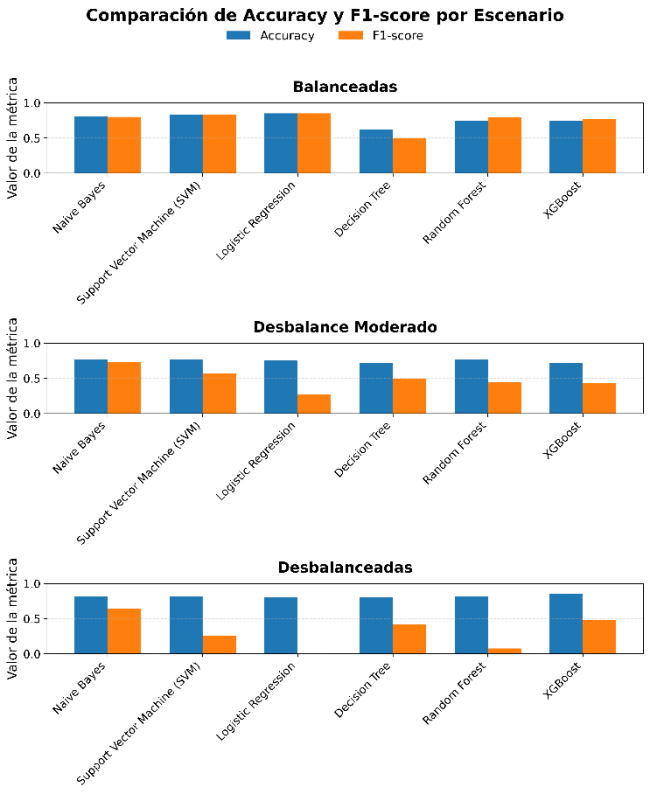


Fig. 2. Comparación de *accuracy* y *f1-score* por escenario

La Figura 2 presenta la comparación entre los valores de exactitud y *f1-score* obtenidos por cada modelo en los tres escenarios experimentales, organizados en grupos de barras que permiten visualizar simultáneamente ambas métricas para cada configuración de distribución de datos, lo que resulta especialmente relevante para examinar la divergencia entre ambos indicadores a medida que el nivel de desequilibrio entre clases se incrementa. En el escenario equilibrado, se observa una correspondencia relativamente estrecha entre la exactitud y el *f1-score* en la mayoría de los modelos, reflejando que bajo condiciones de distribución balanceada ambas métricas ofrecen una caracterización consistente del rendimiento clasificatorio, siendo la *Logistic Regression* un ejemplo paradigmático de este comportamiento al registrar valores prácticamente idénticos en ambas métricas, con 85,1% tanto en exactitud como en *f1-score*.

Sin embargo, a medida que el desequilibrio entre clases se acentúa, emerge una divergencia progresiva y sistemática entre ambos indicadores que se vuelve especialmente pronunciada en el escenario severamente desequilibrado, siendo los casos más ilustrativos de este fenómeno la *Logistic Regression* y *Random Forest*, ya que el primero mantiene una exactitud del 80,6% en dicho escenario mientras su *f1-score* colapsa a 0,0%, en tanto que el segundo alcanza una exactitud del 81,4% con un *f1-score* residual de tan solo 7,7%. Esto pone de manifiesto que, en presencia de desequilibrio severo, la exactitud puede reflejar valores aparentemente satisfactorios como consecuencia de la predicción sistemática de la clase mayoritaria, sin que ello implique una capacidad real de discriminación entre categorías.

Este comportamiento constituye una evidencia empírica contundente de las limitaciones de la exactitud como indicador único de rendimiento en contextos con distribuciones

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.02

asimétricas. En efecto, un modelo que clasifica la totalidad de las instancias como pertenecientes a la clase negativa, la predominante en el *dataset*, puede alcanzar una exactitud global cercana al 81,8%, correspondiente a la proporción natural de dicha clase, sin haber aprendido patrón discriminativo alguno, por lo que el *f1-score*. Por ello, el *recall* mediante su media armónica, resulta sensible a este tipo de comportamiento degenerativo y lo penaliza de forma proporcional, ofreciendo así una valoración más fidedigna de la capacidad real del clasificador.

La importancia de adoptar un enfoque multimétrico en la evaluación de modelos de clasificación es resaltada por los resultados de este análisis, especialmente en tareas donde existe un desequilibrio entre clases. La utilización exclusiva de la exactitud como criterio único de evaluación en estos escenarios puede conducir, en consecuencia, a conclusiones erróneas sobre la calidad real de los modelos, siendo la métrica *f1-score*, en particular, un indicador indispensable para lograr una valoración con mayor rigurosidad a la hora de medir el desempeño clasificatorio de los modelos.

V. Conclusiones

Los hallazgos derivados del presente estudio demuestran de manera inequívoca que el grado de desequilibrio entre clases representa un factor de influencia crítica sobre el rendimiento de los modelos de aprendizaje automático aplicados al análisis de sentimientos en tweets en español. De modo que, a lo largo de las tres configuraciones experimentales evaluadas, se constata que el comportamiento de los algoritmos no únicamente fluctúa en términos de desempeño absoluto, sino también en su capacidad intrínseca de adaptación ante distribuciones de datos progresivamente más asimétricas, poniendo así de manifiesto la complejidad inherente a este tipo de problemas en contextos reales.

En el escenario de distribución equilibrada, los clasificadores de naturaleza lineal, en particular la *Logistic Regression* y el *SVM*, evidenciaron una superioridad manifiesta frente al resto de los algoritmos evaluados, logrando un balance adecuado entre precisión y *recall* que se tradujo en valores elevados de *f1-score*. Esto sugiere que la arquitectura de estos modelos resulta especialmente apropiada cuando las categorías se encuentran representadas de forma proporcional en el conjunto de datos, hallazgo que, además, guarda consistencia con lo reportado en la literatura especializada, donde los clasificadores lineales tienden a destacar en tareas de clasificación textual bajo condiciones ideales de distribución [33].

No obstante, una degradación considerable en el rendimiento de estos modelos es desencadenada por la introducción de un desequilibrio moderado, siendo el caso más ilustrativo de esta vulnerabilidad la *Logistic Regression*, pues, a pesar de preservar valores elevados de precisión, es su *recall* el que experimenta una reducción drástica. En consecuencia, esto revela una tendencia progresiva a favorecer la clase dominante en detrimento de la capacidad para recuperar instancias de la categoría minoritaria, situación que se agudiza de forma extrema en el escenario severamente desequilibrado, donde en un colapso clasificatorio total incurre el modelo, y evidencia una incapacidad absoluta para identificar ejemplos positivos. A su vez, de manera análoga, contracciones pronunciadas en su desempeño son registradas por el *SVM* y *Random Forest*, lo que corrobora, por ende, su elevada sensibilidad ante las perturbaciones introducidas por el desequilibrio entre clases.

En contraposición, como el algoritmo de mayor estabilidad a lo largo del espectro de escenarios evaluados emerge *Naive Bayes*, muestra que la degradación de su *f1-score* sigue una trayectoria progresiva y notablemente más contenida en comparación con los restantes modelos, lo que sugiere, en consecuencia, que su fundamento probabilístico le proporciona una resiliencia inherente para sostener una capacidad clasificatoria consistente incluso ante distribuciones adversas, Además, pone de relieve su valor práctico en entornos reales donde el desequilibrio entre categorías constituye una condición estructural de los datos más que una anomalía excepcional.

A lo largo del análisis, se identifica un aspecto de relevancia metodológica en la relación dinámica entre las métricas de evaluación empleadas. Se observa que, en condiciones de distribución equilibrada, la exactitud y el *f1-score* convergen hacia valores similares y, por tanto, ofrecen representaciones concordantes del rendimiento de los modelos. Sin embargo, a medida que el desequilibrio entre clases se intensifica, emerge, en consecuencia, una divergencia sistemática y creciente entre ambos indicadores. De forma particularmente llamativa, varios modelos preservan valores de exactitud aparentemente satisfactorios al mismo tiempo que su *f1-score* se desploma hacia valores mínimos. Este fenómeno expone la naturaleza engañosa de la exactitud como criterio único de evaluación en contextos desequilibrados, puesto que este indicador puede inflarse artificialmente por la predicción sistemática de la clase mayoritaria sin que ello implique capacidad discriminativa real alguna.

En este sentido, la métrica *f1-score* se posiciona como el indicador de referencia para la evaluación del desempeño en presencia de desequilibrio entre clases, al integrar simultáneamente la precisión y el *recall* . Además, penaliza de forma proporcional los errores asociados a la clase minoritaria, lo cual refuerza la necesidad de adoptar estrategias de evaluación multimétrica en las tareas de clasificación aplicadas a problemas reales, donde la uniformidad en la distribución de los datos suele ser más la excepción que la norma.

El análisis comparativo transversal entre escenarios permite establecer que no existe un algoritmo universalmente óptimo para todas las condiciones de distribución, sino que la superioridad relativa de cada modelo se encuentra fuertemente condicionada por las características propias de la distribución de los datos. En este sentido, como contribución principal del presente estudio se destaca la evaluación empírica y estructurada del comportamiento de múltiples modelos de aprendizaje automático ante distintos niveles de desequilibrio de clases en un contexto aplicado de análisis de sentimientos sobre datos provenientes de redes sociales. Todo esto evidencia que *Naive Bayes* y *XGBoost* representan opciones preferentes en escenarios con distribuciones asimétricas, mientras que un rendimiento superior bajo condiciones de equilibrio entre categorías es ofrecido por los modelos lineales. Estos resultados constituyen un aporte de valor práctico para investigadores y profesionales que enfrentan problemas de clasificación en dominios donde el desequilibrio de clases es una realidad estructural ineludible.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.02

A pesar de los aportes descritos, el presente estudio no está exento de limitaciones, puesto que el corpus utilizado se circunscribe a un dominio temático específico y a un período temporal determinado, lo que podría restringir la generalización de los hallazgos a otros contextos discursivos o a datos provenientes de distintas épocas o plataformas. Asimismo, el análisis se realizó sin la aplicación de técnicas de balanceo de clases como *SMOTE* o *undersampling*, ni de estrategias de preprocesamiento avanzado, aspectos que potencialmente habrían permitido mitigar el impacto del desequilibrio sobre el rendimiento de los modelos, en este sentido, investigaciones futuras podrían explorar el efecto combinado de dichas técnicas de balanceo junto con modelos de lenguaje preentrenados como *BERT* o *RoBERTa* en español, así como ampliar el análisis a corpus multidominio y multiclase, con el objetivo de obtener una comprensión más robusta y generalizable del comportamiento de los clasificadores ante distribuciones asimétricas en tareas de análisis de sentimientos.

REFERENCES

[1] S. Giménez, "Redes Sociales, estado actual y tendencias 2023 OBSbusiness.school," 2023.

[2] A. Albladi, M. Islam, and C. Seals, "Sentiment Analysis of Twitter Data Using NLP Models: A Comprehensive Review," IEEE Access, vol. 13, pp. 30444–30468, 2025, doi: 10.1109/ACCESS.2025.3541494.

[3] J. J. Moreno and R. M. Nicolás, "Evaluación de la percepción ciudadana en la red social X mediante técnicas de minería y analítica de datos para el fortalecimiento institucional de la Secretaría Distrital de Hacienda," Oct. 2025, Accessed: Mar. 16, 2026. [Online]. Available: http://repository.unad.edu.co/handle/10596/78213

[4] I. J. Girón, "Análisis comparativo de modelos de aprendizaje supervisado para el reconocimiento de emociones en texto," 2025, Accessed: Mar. 16, 2026. [Online]. Available: https://uvadoc.uva.es/handle/10324/79497

[5] M. E. Jonatan, "Clasificación de datos desbalanceados," May 2022, Accessed: Mar. 16, 2026. [Online]. Available: http://sedici.unlp.edu.ar/handle/10915/147410

[6] J. J. Campo Yepes, "Evaluación de métricas en modelos predictivos de clasificación en machine learning." Accessed: Mar. 16, 2026. [Online]. Available: https://repository.universidadean.edu.co/entities/publication/81ddd265-219d-480f-8422-1e5486ad0c75

[7] E. M. Porras, A. R. Fernández, L. P. S. Bastos, and J. A. D. González, "Uso de Procesamiento de Lenguaje Natural para procesar respuestas abiertas de una encuesta de Opinión Pública," Revista Latinoamericana de Metodología de la Investigación Social: ReLMIS, no. 29, pp. 51–67, 2025.

[8] J. E. Lozano González, "Revisión sistemática sobre el análisis de sentimientos en interacciones por chat en videojuegos," Feb. 2025, Accessed: Mar. 16, 2026. [Online]. Available: http://repository.unad.edu.co/handle/10596/67163

[9] N. M. Sanchez Posada, "Optimización del marketing digital: revisión sistemática de técnicas de Big Data y monitoreo de redes sociales," Aug. 2025, Accessed: Mar. 16, 2026. [Online]. Available: http://repository.unad.edu.co/handle/10596/73389

[10] L. Rivadeneira, "Análisis del comportamiento de decisión usando - ProQuest," Revista Ibérica de Sistemas e Tecnologias de Informação, 2023.

[11] E. G. Mita Arancibia, "Revisión sistemática sobre análisis de datos en tiempo real: Herramientas para tomar decisiones estratégicas," Panel - Revista de Administración, Jul. 2024.

[12] A. D. Jiménez Alfaro and J. V. Díaz Ospina, "Revisión sistemática de literatura: Técnicas de aprendizaje automático (Machine Learning)," Cuaderno activa, vol. 13, no. 1, pp. 113–121, 2021.

[13] A. Cevallos-Culqui, C. Pons, and G. Rodriguez, "Semi-supervised learning models for document classification: A systematic review and meta-analysis," Inteligencia Artificial, vol. 26, no. 72, pp. 81–111, Jun. 2023, doi: 10.4114/intartif.vol26iss72pp81-111.

[14] M. A. Hernández Castañeda and M. F. Forero Dorado, "Uso de algoritmos Machine Learning en la clasificación de objetos astronómicos: una revisión sistemática," UNAD, Dec. 2024.

[15] J. L. Romero Ibarra, "Análisis integral de algoritmos de clasificación en aprendizaje automático: perspectivas, comparaciones y aplicaciones," Serie Científica de la Universidad de las Ciencias Informáticas, vol. 18, no. 1, pp. 283–304, 2025.

[16] A. F. Ruiz Delgado, "Revisión sistemática de modelos de machine learning y deep learning aplicados a la detección temprana de depresión en redes sociales," UNAD, Dec. 2025.

[17] R. Tobar-Díaz, Y. Gao, J. F. Mas, and V. H. Cambrón-Sandoval, "Classification of land use and land cover through machine learning algorithms: a literature review," Revista de Teledetección, vol. 2023, no. 62, pp. 1–19, Jul. 2023, doi: 10.4995/raet.2023.19014.

[18] V. H. Bustamante Morán, C. E. Quiroz Calle, J. R. Oquendo Silva, and J. R. Oquendo Silva, "Inteligencia artificial en el diagnóstico diferencial de patologías de tejidos blandos orales," RECIAMUC, vol. 9, no. 4, pp. 472–492, Dec. 2025, doi: 10.26820/reciamuc/9.(4).diciembre.2025.472-492.

[19] L. J. Montesdeoca Espinoza, S. J. Zambrano Rojas, V. J. Pinargote-Bravo, and L. Cedeño-Valarezo, "Redes generativas para balanceo de datos en imágenes agrícolas: una revisión sistemática de la literatura," Revista Científica de Informática ENCRIPTAR, vol. 8, no. 16, pp. 153–168, Oct. 2025, doi: 10.56124/encriptar.v8i16.008.

[20] H. P. Segovia Granda, "Revisión sistemática y análisis de metodologías que utilizan técnicas de minería de datos y aprendizaje automático para detección del trolling en las redes sociales," 2022.

[21] C. Huamañí Ninahuanca, C. J. Quintana-Castro, and N. E. Tovar Soto, "Modelo predictivo sobre pérdida de beca por motivos académicos en beneficiarios de Beca 18," 2025.

[22] G. Poquechoque Foronda and C. W. Pacheco Lora, "Evaluación de la experiencia de usuario ante interfaces web de software de gestión a través del análisis de emociones," Revista Ciencia y Tecnología Digital, vol. 1, no. 1, pp. 1–20, Oct. 2025.

[23] F. E. Garza, Y. M. Ramírez, A. R. Noriega, and I. N. Á. Sánchez, "Una revisión sistemática sobre la precisión de modelos de aprendizaje automático aplicados a la tasación de bienes raíces," RITI, vol. 12, no. 28, pp. 4–16, 2024, doi: 10.36825/RITI.12.28.002.

[24] J. E. Pino Cotillo, "Comparación de modelos de machine learning para la predicción temprana de diabetes mellitus tipo 2," UNAD, Dec. 2025.

[25] O. Ali, W. Abdelbaki, A. Shrestha, E. Elbasi, M. A. A. Alryalat, and Y. K. Dwivedi, "A systematic literature review of artificial intelligence in the healthcare sector," Journal of Innovation & Knowledge, vol. 8, no. 1, p. 100333, Jan. 2023, doi: 10.1016/j.jik.2023.100333.

[26] S. A. Montero Franco, "Revisión Teórica De Las Redes Neuronales Profundas Para La Detección De Malware," UNAD, Aug. 2025.

[27] J. I. Valdés Espinoza, "Clasificación automatizada de actividad cerebral anormal en pacientes neurocríticos para mejorar capacidad diagnóstica," 2022, doi: 10.58011/X5MP-TG04.

[28] E. Cruz, M. González, and J. C. Rangel, "Técnicas de machine learning aplicadas a la evaluación del rendimiento y a la predicción de la deserción de estudiantes universitarios, una revisión," Prisma Tecnológico, vol. 13, no. 1, pp. 77–87, Feb. 2022, doi: 10.33412/pri.v13.1.3039.

[29] R. S. Cedeño Menéndez, J. A. León Alarcón, and J. H. Franco Cantos, "Análisis de Sentimientos en la Red Social 'X', Percepción Pública sobre el Presidente del Ecuador, Daniel Noboa (noviembre 2023 - abril 2024)," Latin-American Journal of Computing, vol. 12, no. 2, pp. 40–48, Jul. 2025, doi: 10.33333/lajc.vol12n2.03.

[30] V. A. García Meza and S. H. Lázaro Barrera, "Revisión teórica de modelos de Machine Learning para la predicción del comportamiento de pago en clientes gestionados desde contact-center, sector cobranzas," UNAD, Dec. 2025.

[31] J. Urrego Piedrahita and J. J. Acosta Jiménez, "Machine Learning aplicado a la predicción de pacientes en EPS: una revisión de literatura," Cuaderno activa, vol. 16, no. 1, May 2024, doi: 10.53995/20278101.1574.

[32] H. Ali, M. Najib, M. Salleh, R. Saedudin, K. Hussain, and M. Faheem Mushtaq, "Imbalance class problems in data mining: a review," Indonesian Journal of Electrical Engineering and Computer Science, vol. 14, no. 3, pp. 1560–1571, 2019, doi: 10.11591/ijeecs.v14.i3.pp1560-1571.

[33] L. U. Hidalgo Vargas, J. A. León Borges, J. C. Ramírez Pacheco, H. Toral Cruz, T. G. Makita Balcorta, and I. Osuna Galán, "Análisis de

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.02

Revisión Sistemática de la Aplicación de Algoritmos de Aprendizaje Automático en Sistemas de Detección de Intrusión en Internet de las Cosas para Ciudades Inteligentes," Ciencia Latina Revista Científica Multidisciplinar, vol. 8, no. 6, pp. 11500–11517, Feb. 2024, doi: 10.37811/cl_rcm.v8i6.15929.

Declaración de Uso de herramientas de inteligencia artificial

Durante la preparación de este manuscrito, por los autores se declara que una herramienta basada en inteligencia artificial fue utilizada exclusivamente con fines de traducción y mejora de la redacción, siendo empleada para optimizar la calidad lingüística y la legibilidad del texto, mientras que, en contraste, el contenido científico, el análisis de datos, las interpretaciones y las conclusiones son responsabilidad íntegra de los autores.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.02

AUTHORS

Roly Steeven Cedeño Menéndez

Ingeniero en Sistemas de Información por la Universidad Técnica de Manabí y Magíster en Sistemas de Información con mención en Data Science por la Pontificia Universidad Católica del Ecuador. Su formación académica y experiencia profesional se enfocan en el análisis de datos, el aprendizaje automático y la aplicación de técnicas avanzadas para la extracción de conocimiento a partir de grandes volúmenes de información. Actualmente se desempeña como técnico docente en la Universidad Técnica de Manabí y cuenta con un año de experiencia adicional como docente en modalidad online.

Ha participado en proyectos de investigación vinculados a la ciencia de datos, destacando su trabajo de tesis de posgrado titulado "Análisis de sentimientos utilizando la red social X (Twitter) para medir el nivel de aceptación del nuevo presidente del Ecuador, Daniel Noboa (noviembre 2023 - abril 2024)". También cuenta con dos artículos académicos publicados. Sus áreas de interés incluyen la inteligencia artificial, la minería de datos y el desarrollo de soluciones basadas en ciencia de datos. Sus objetivos profesionales actuales se centran en mejorar continuamente como docente y consolidarse como investigador en el área, contribuyendo con nuevas publicaciones científicas.

José Alberto León Alarcón

José León Alarcón es un profesional especializado en Ciencia de Datos, posee un máster en Sistemas de Información con mención en Data Science por la Pontificia Universidad Católica del Ecuador (PUCE Quito). Su formación académica se complementa con una sólida experiencia en el ámbito de la inteligencia artificial, especialmente en el aprendizaje automático (machine learning) y el aprendizaje profundo (deep learning). A lo largo de su trayectoria profesional, se ha enfocado en el análisis de imágenes médicas, contribuyendo al desarrollo de modelos capaces de apoyar el diagnóstico clínico mediante técnicas avanzadas de procesamiento de imágenes. Además, ha trabajado en la extracción y análisis de información a partir de datos complejos, aplicando metodologías estadísticas y herramientas computacionales modernas. Sus áreas de interés incluyen la inteligencia artificial, el análisis predictivo y el desarrollo de soluciones innovadoras que permitan transformar grandes volúmenes de datos en conocimiento útil para la toma de decisiones. Se caracteriza por su compromiso con la investigación aplicada y el desarrollo tecnológico orientado a resolver problemas reales.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026

AUTHORS

Jandry Franco Cantos

Ingeniero en Sistemas de Información con una Maestría en Ingeniería en Sistemas de Información, mención en Data Science. Ha formado parte de diversos proyectos enfocados en el desarrollo de software e implementación de soluciones basadas en inteligencia artificial, aplicadas al análisis de datos, la optimización de procesos y la automatización de tareas.

Actualmente se desempeña como docente universitario en la Universidad Técnica de Manabí, Ecuador, donde combina la formación académica con la investigación aplicada. Sus principales áreas de interés incluyen la inteligencia artificial, el aprendizaje automático, la visualización de datos y la ciencia de datos orientada a la toma de decisiones.

Cuenta con experiencia en la integración de herramientas tecnológicas en entornos educativos y productivos, participando activamente en iniciativas interdisciplinarias que promueven la innovación tecnológica con impacto real. Su enfoque profesional se basa en el desarrollo de soluciones prácticas y eficientes, alineadas con los avances actuales en ciencia y tecnología.

Comprometido con la formación de nuevas generaciones de profesionales, busca contribuir al avance del conocimiento científico y al desarrollo de tecnologías sostenibles que respondan a las necesidades actuales de la sociedad.

R. S. Cedeño Menéndez, J. A. León Alarcón, and J. H. Franco Cantos, "Evaluación del rendimiento de modelos de Machine Learning para el análisis de sentimientos en tweets en español", Latin-American Journal of Computing (LAJC), vol. 13, no. 2, 2026.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026

Morphological classification of hematophagous Diptera with Convolutional Neural Networks: A mapping of literature

ARTICLE HISTORY
Received 14 October 2025
Accepted 23 February 2026
Published 7 July 2026

Benjamín Paulino Mendoza Contreras
Veracruzana University
Faculty of Statistics and Informatics
Xalapa, Veracruz
benjaminpaulinom6@gmail.com
ORCID: 0009-0000-3491-6234

Emmanuel Morales García
Veracruzana University
Faculty of Statistics and Informatics
Xalapa, Veracruz
emmorales@uv.mx
ORCID: 0000-0002-6837-9227

Cecilia Cruz López
Veracruzana University
Faculty of Statistics and Informatics
Xalapa, Veracruz
ceccruz@uv.mx
ORCID: 0000-0002-9156-5669

Luis Enrique Gomez Medina
Veracruzana University
Institute for Research and Higher Studies in Administrative Sciences
Xalapa, Veracruz
luisgomez04@uv.mx
ORCID: 0009-0009-1324-389X

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026

Morphological classification of hematophagous Diptera with Convolutional Neural Networks: A mapping of literature

Benjamín Paulino Mendoza Contreras
Veracruzana University
Faculty of Statistics and Informatics
Xalapa, Veracruz
benjaminpaulinom6@gmail.com

Emmanuel Morales García
Veracruzana University
Faculty of Statistics and Informatics
Xalapa, Veracruz
emmorales@uv.mx

Cecilia Cruz López
Veracruzana University
Faculty of Statistics and Informatics
Xalapa, Veracruz
ceccruz@uv.mx

Luis Enrique Gomez Medina
Veracruzana University
Institute for Research and Higher Studies in Administrative Sciences
Xalapa, Veracruz
luisgomez04@uv.mx

***Abstract*—** This review analyzes studies that primarily address the morphological classification of hematophagous Diptera, with limited mention of other insects. These networks have become increasingly important in morphological analysis through the accurate and efficient automatic identification of species, surpassing even traditional methods based on human observation. The main architectures used, such as VGG-16, YOLOv5, Faster R-CNN, Mask R-CNN, ResNet, and Swin Transformer-L are reviewed, highlighting their applications in the detection and identification of different anatomical parts. Common limitations are also mentioned, such as the need for large volumes of classified data and variability in image quality. Finally, current trends have been identified that point to the development of more robust hybrid models capable of recognizing new species and improving accuracy under real-world conditions. This literature mapping provides greater certainty and evidence regarding the most important identification methods in the field of entomology. These findings highlight the gap in literature related to the availability of public data, parameters used, data volume, image quality, and model evaluation, providing a solid foundation to guide future research in the field of entomology.

Keywords*—Entomology, Organism Classification, Deep Learning, Species, Identification, Morphology*

I. Introduction

Image-based classification of hematophagous Diptera is fundamental and has evolved thanks to the various Convolutional Neural Networks (CNNs) used for image recognition. These networks allow for the analysis of visual characteristics (morphology) of species at a higher level and more quickly, even under varying image conditions [1].

Classifying hematophagous Diptera from digital images allows for faster, more accurate, and scalable identification than traditional methods based on human observation. These networks are capable of automatically extracting complex visual features (shapes, textures, or patterns) that determine the differences between species, without requiring the researcher to manually define the important characteristics [1].

One of the main advantages of CNNs is their ability to process large volumes of data from different image capture devices. This represents a significant change in field data collection, as it automates repetitive tasks and reduces the human workload, allowing experts and researchers to achieve greater accuracy.

Some authors have developed hierarchical architectures that incorporate taxonomic relationships between genera and species within the model itself, improving performance and reducing errors when classifying at more specific taxonomic levels [2],[3]. This type of innovation is important for classifying morphologically similar organisms, where visual differences may be minimal.

Convolutional neural networks (CNNs) are one of the most influential innovations in the field of deep learning, due to their ability to automatically and efficiently process, analyze, and classify visual data. Their main advantage lies in their ability to extract hierarchical features directly from images, reducing the need for human intervention in selecting relevant features [4].

In the scientific and technological fields, CNNs have demonstrated exceptional performance in tasks such as facial recognition, medical imaging, organism classification, object detection, and autonomous driving. Their structure, based on convolutional, clustering, and fully connected layers, allows for the identification of complex patterns [5],[6].

The objective of this literature mapping is to show the trends, methodological approaches, data types, and CNN architecture used in the morphological classification of Diptera reported in recent research. Although morphological classification includes various biological groups, recent studies show a significant concentration on hematophagous Diptera, due to their relevance to public health issues. Therefore, this research focuses on this group.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.03

II. Methodology

This study employs an exploratory and descriptive literature review approach, aiming to identify trends, methodologies, convolutional neural network (CNN) architectures, data types used, and gaps in the literature regarding morphological classification of dipterans using images. This type of review seeks to provide a structured overview of the current state of the art. Fig. 1 shows the general phases of the mapping review process followed in this study.

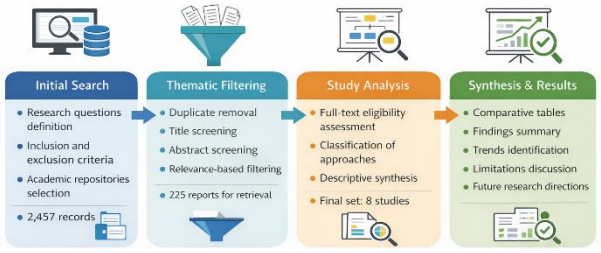


Fig. 1. Article selection process.

A. Research Questions

RQ1. What pre-trained convolutional neural network architectures have been used to detect morphology in Diptera?

RQ2. What CNN-based approaches analyze Diptera morphology?

RQ3. What types of results are reported in studies on wing morphology?

RQ4. What computational approaches have been proposed for detecting morphological patterns in Diptera?

B. Inclusion and exclusion criteria

TABLE I. Selection criteria for primary studies

Category	Inclusion	Exclusion
Type of research	Practical research on image classification methods using convolutional neural networks	Non-primary studies: literature review.
Publication year interval	Articles published from 2020 to 2025 to ensure current relevance	Studies published before 2020.
Language	Articles in English	Articles in Spanish or another language.
Search engines	Primary Search Engines (Publishers)	Search engines of dubious scientific quality
Applications	Emphasis on medical imaging and other areas of study	
Thematic relevance	Research on the classification of diptera	Investigations that divert attention from the main topic

C. Search Strategy

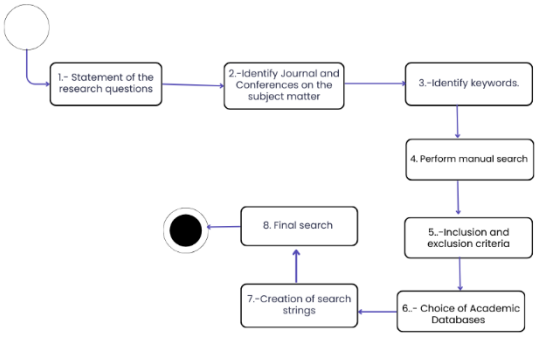


Fig. 2. Distribution of selected articles by digital library.

For this study, a search strategy was employed that was designed to achieve broad and specific coverage of the relevant literature, following a descriptive process (Fig. 2). The following elements were considered:

1. Repositories used: ACM Digital Library, IEEE Xplore, SpringerLink, and ScienceDirect.
2. Year range: Publications between 2020 and 2025, with the aim of including recent work.
3. Language: Only publications in English.
4. Keywords: Terms related to CNNs, image classification, Diptera, and public health.

To address the variability of terms present in the literature, a set of keywords and synonyms were proposed, which were combined using Boolean operators. The general search string used was:

"convolutional neural network" OR CNN OR "deep learning" OR "computer vision") AND ("image classification" OR "image recognition" OR "object detection") AND (morphology OR "morphological identification" OR "morphometric analysis") AND (Diptera OR mosquito OR mosquitoes OR Culicidae OR "hematophagous insects") AND (wing OR winVuelgs OR "wing morphology" OR "body morphology"

This string is adapted to the specific syntax of each repository used.

The annual trend in publications in the selected subsample was analyzed, as shown in Fig. 3. The frequency analysis by year for the period 2020-2025 shows a growing trend in the publication of articles that use Convolutional Neural Networks (CNN) in the analyzed repositories.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.03

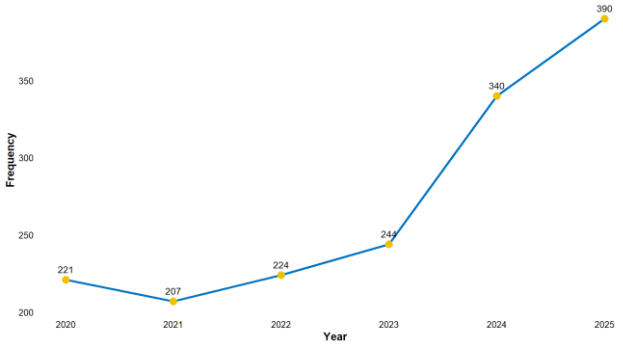


Fig. 3. Annual trend in publication frequency (2020-2025) of the subsample of articles selected from the four repositories

In addition to the trend analysis, the authors with the highest number of scientific publications were identified using CNNs to classify organisms in the four academic repositories shown in Fig. 4. The word cloud in Fig. 4 reveals a clear concentration of publications by Asian authors, particularly those with the surnames Zhang, Liu, Li, and Wang.

Fig. 4. Word cloud of the most frequent authors in the publication of articles in the four repositories (2020-2025)

D. Study Selection Process

The search conducted across the selected academic repositories yielded a total of 2,457 records (Table II). To ensure the relevance of the studies included in this mapping review, a multi-stage filtering process was applied following the PRISMA 2020 guidelines (Fig. 5). First, duplicate records were removed, resulting in the exclusion of 412 articles, leaving 2,045 unique records for the screening phase.

Second, a thematic filtering stage was performed based on the analysis of titles, abstracts, and keywords, which led to the exclusion of 1,820 records that were not aligned with the objectives of this study. As a result, 225 articles were considered potentially relevant and were retrieved for full-text evaluation.

Third, during the eligibility assessment, the full texts of the 225 remaining studies were examined in detail. At this stage, 217 studies were excluded because they did not meet the inclusion criteria. The main reasons for exclusion included studies not focused on morphological classification using images, studies that did not employ deep learning methods, articles lacking experimental evaluation or performance metrics, studies based on non-image data, and research outside the scope of insect identification. After applying these criteria, a final set of eight studies was selected for qualitative analysis. These studies were analyzed to identify the architecture, datasets, and methodological trends used in automated morphological classification of insects using artificial intelligence.

TABLE II. NUMBER OF ARTICLES RECOVERED FROM SELECTED ACADEMIC REPOSITORIES (2020–2025)

Repository	Total, number of items
ACM	165
IEEE Xplore	116
SpringerLink	397
ScienceDirect	1779

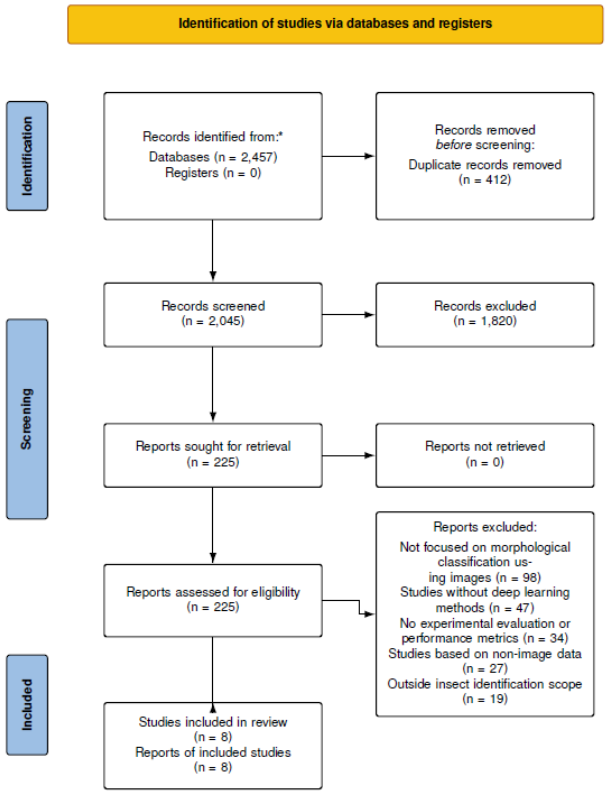


Fig. 5. Study filtering process

E. Descriptive analysis of the studies

Table III below shows information about the selected studies, allowing you to view key information about each one. It should be noted that no quantitative comparison is made.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.03

TABLE III. ARTICLES DATA MATRIX

Author(s)	Year	Organism	Analyzed part	Dataset	Task type	Metrics
Adhane, G., Dehshibi, M. M., & Masip, D. [1]	2021	Mosquitoes (A. albopictus)	Body (legs, abdomen)	Public	Classification	94.61% Accuracy
Minakshi, M. et al. [21]	2020	Mosquitoes (9 species)	Thorax, wings, abdomen	Private	Detection/Feature extraction	95% Accuracy, 60% mAP
Sauer, F. G. et al. [22]	2024	Mosquitoes (Aedes)	Wings	Private	Classification	99% F1-score
Nolte, K. et al. [23]	2025	Mosquitoes (4 species)	Wings vs body	Private	Classification	87.6% (Wings) / 78.9% (Body)
Cannet, A. et al. [25]	2023	Mosquitoes (Aedes genus)	Interference patterns (wings)	Private	Classification	95% Accuracy
Zhao, D. et al. [28]	2022	Mosquitoes (17 species)	Whole body	Private	Classification	99.04% Accuracy
Lee, S., Kim, H., & Cho, B.-K. [29]	2023	Mosquitoes (11 species)	Whole body	Mixed	Detection	97.1% F1-score
Goodwin, A. et al. [30]	2021	Wildlife species	Body	Private	Multilevel	97.04% Accuracy (Known classes)

III. CONCEPTUAL FOUNDATIONS

A Convolutional Neural Network (CNN) is a type of deep learning model designed primarily to process data with a grid-like structure, such as images or spatial and temporal signals. CNNs are inspired by the functioning of the visual cortex of the human brain, which responds selectively to visual patterns such as edges, textures, and shapes [4].

It is a machine learning system that mimics human visual perception, capable of learning hierarchical patterns and complex representations from large volumes of visual data. Thanks to their generalization capacity and efficiency, CNNs have become an essential tool in modern artificial intelligence and in various areas of application [4].

The fundamental principle of a CNN is the convolution operation, whereby the model applies filters (also called kernels) to images to automatically extract relevant features at different levels of abstraction. In the first layers, simple features such as edges or colors are detected, while in the deeper layers, more complex shapes such as complete objects are recognized [4].

These networks have demonstrated robust performance in computer vision tasks such as facial recognition, medical image diagnosis, organism classification, object detection, and autonomous driving. Unlike traditional methods, CNNs do not require an expert to manually define visual features, as they learn directly from the data, which increases accuracy and reduces human bias. In addition, CNNs achieve very high levels of accuracy in visual recognition tasks, outperforming image classification or object detection. Thanks to the use of graphics processing units (GPUs), these networks can handle large volumes of visual information in a short time, making them efficient and scalable [7].

However, CNNs also have limitations. They require large amounts of labeled data to achieve adequate performance, as well as high computational costs, which can limit their application in contexts with limited data and resources. Furthermore, when the dataset is small or lacks variety, the model may overfit, i.e., learn specific patterns from the training that do not generalize correctly to new data [3].

Another limitation relates to the lack of interpretability of the results. CNNs are considered "black box" models because it is not always possible to know exactly how they make decisions, which creates uncertainty in areas where explaining the process is as important as the prediction, such as in medicine or biology. Similarly, their performance can be affected by variability in external conditions, such as camera angle, the image capture device used, or the background of the images. Finally, if the training data is biased or imbalanced between classes, the model may reproduce those same biases in its results, affecting the accuracy of the predictions [8].

CNNs are a fundamental tool for automated image processing and analysis, with applications in numerous fields of knowledge. They have transformed the analysis of biological images by offering an automated, efficient, and accurate way to process large volumes of visual information. In recent years, these architectures have become an important tool in computational biology, ecology, and digital taxonomy, enabling species identification, organism counting, and morphological characterization from photographs or video sequences. Likewise, they have shown outstanding results in the identification of animal species captured by camera traps or drones, achieving greater accuracy than experts [9].

Another limitation in the analysis of biological images is the scarcity of labeled data, since obtaining high-quality images for each species or taxonomic group is costly and requires expert knowledge. This problem becomes more relevant when there are unbalanced classes, causing the model to favor the classes with more images and reduce the accuracy of the minority classes [10].

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.03

IV. LITERATURE REVIEW AND DESCRIPTIVE SYNTHESIS OF THE RESEARCH QUESTIONS

A. RQ1. What pretrained convolutional neural network architectures have been used to detect morphology in diptera?

It is designed for large-scale image classification and recognition. It consists of 16 weighted layers (13 convolutional and 3 dense). Its key architectural principle is the exclusive use of 3×3 convolutional filters stacked in blocks, followed by Max Pooling layers (2×2). The process begins with an input image (224×224 RGB), from which features are extracted to be finally classified by the dense layers with Softmax activation [11].

Faster R-CNN

It is a two-stage object detection model based on Convolutional Neural Networks (CNN):

1. Region Proposal Network (RPN): Uses anchors with various scales and aspect ratios to propose regions, classifying them and adjusting their coordinates (regression).
2. Detection (Fast R-CNN): It uses ROI Pooling/Align to standardize the proposals, then performs the final classification of the object and the final refinement of the bounding box.

The parameters include the choice of backbone, the configuration of the anchors, and the use of Soft L1 Loss for box regression [12].

YOLOv5

It is a single-stage real-time object detector with an architecture divided into three parts:

1. Backbone: Uses CSPDarknet53 for efficient feature extraction.
2. Neck: Employs SPPF and PANet to fuse and enhance features at different scales, which is key to its high performance.
3. Head: This is the final layer that performs simultaneous prediction of class, objectivity, and bounding box coordinates.

The parameters are defined by model variants that control the depth and width of the network, using the CIoU loss function (for localization) and BCE (for classification) [13].

Darknet

It is the convolutional architecture that serves as the backbone for the YOLO algorithm, optimized for real-time object detection. The Darknet process (especially in versions such as Darknet-53) is based on a fully convolutional architecture that incorporates residual connections (like ResNet) to increase depth. Its key parameters include the use of 1×1 layers for dimensionality reduction and the Leaky ReLU (or Mish) activation function to improve gradient flow, allowing YOLO to predict bounding boxes and classes in a single pass through the network [14].

ResNet101 DC-5

It is a 101-layer Deep Residual Neural Network that uses residual connections to prevent accuracy degradation in deep learning model training.

It uses ResNet base architecture with bottleneck blocks. The DC-5 feature involves replacing the pooling layers in Stage 5 with dilated convolutions. Its main application is image classification, although it can be optimized for detection and semantic segmentation tasks [15].

ResNeXt101

It is an architecture that extends ResNet to improve accuracy by introducing cardinality as a new dimension of scalability. It is based on Transformation Aggregation. Each residual block divides it into multiple parallel and identical branches that process different aspects of the input before merging. Its main parameter is Cardinality (C), which is the number of independent groups or branches, usually using a value of C=32 [15].

ResNet18

It is the smallest and most efficient version of the Residual Neural Network family, designed for image classification and to serve as a fast backbone. Its most important process is the use of residual connections so that the gradient flows directly through its 18 layers with weights, solving the problem of accuracy degradation in deep networks. Unlike larger versions, it uses basic residual blocks (two 3×3 layers) and has approximately 11.7 million parameters, which gives it speed and efficiency in training [16].

RetinaNet

It is a single-stage image detector designed to achieve high accuracy in dense object detection. Its architecture combines a ResNet backbone with a Feature Pyramid Network (FPN) to process information at multiple scales. Its key process and characteristic parameter are Focal Loss, a modified loss function that resolves the extreme imbalance between positive and negative examples by heavily weighing difficult (misclassified) samples and reducing the weight of easy (background) samples, enabling effective training [17].

Mask R-CNN

It is an advanced model for instance detection and segmentation that extends Faster R-CNN, which detects objects and generates a precise pixel mask for each one. The architecture uses a ResNet-101 backbone combined with a Feature Pyramid Network (FPN) to create a multiscale feature pyramid (neck).

Its process is ROI Align, a parameter that replaces ROI Pooling to ensure precise alignment of the characteristics of the Regions of Interest (ROI) in order to achieve accuracy in the mask. It operates with three parallel heads (classification, box regression, and an FCN for the mask) for application in Instance Segmentation [18].

Swin Transformer-L

It is a hierarchical Vision Transformer architecture with high capacity, designed to be a backbone for achieving high performance in tasks such as object detection and segmentation. Its hierarchical architecture mimics CNNs and its process is Shifted Window Attention (SW-MSA). This mechanism limits the self-attention calculation to local windows, solving the quadratic complexity of traditional Transformers and allowing communication between neighboring windows, resulting in a network with approximately 197 million parameters and superior efficiency and accuracy in vision applications [19].

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.03

MobileNet

It is a lightweight CNN architecture for detection, classification, and segmentation on mobile and edge devices with limited resources. Its process and characteristic feature are the use of Depth Separable Convolutions (DSC), which divide the convolution operation into two steps (depth and point), significantly reducing the number of parameters and the computational load compared to standard convolutions. Its scaling parameters (Width Factor and Resolution Factor) allow the model to be adjusted for different latency and power constraints [20].

B. RQ2. What CNN based approaches analyze Diptera morphology?

Automatic identification of mosquito species has become a central field of research, focusing on the use of convolutional neural networks (CNN) and other deep learning techniques to overcome the challenges of manual classification. Recent studies demonstrate the feasibility and high accuracy of these methods by evaluating the effectiveness of different model architectures and specific anatomical parts used for identification [1].

A key approach focuses on the use of citizen science images, which introduce variability and field conditions. [1] proposed a method for the automated classification of Aedes albopictus mosquitoes using the VGG16 architecture. Unlike other studies based on laboratory images, this model was trained with a large dataset of field images collected by volunteers through the citizen science initiative "Mosquito Alert." Despite the way these images were obtained, the model's VGG16 architecture achieved a test accuracy of 94.61%, demonstrating the feasibility of using neural networks to classify mosquitoes. They applied the Grad-CAM algorithm, and this analysis revealed that the model focuses on the white stripes located on the mosquito's legs, abdomen, and thorax, the same characteristics that entomologists use for identification. It was found that classification errors were directly related to poor image quality, such as lack of clarity, occlusion, or damage to key parts of the mosquito's body. As a result, images of non-tiger mosquitoes with morphological similarities could be misclassified, highlighting the importance of image quality for model accuracy.

In the field of anatomical feature extraction, [21] developed a Mask R-CNN-based framework to automatically detect and extract anatomical components of mosquitoes (thorax, wings, abdomen, and legs) from images obtained with smartphones. They used 1,600 images of nine species for training and validation, and evaluated performance with metrics such as Precision, Recall, IoU, and mAP. The system used ResNet-101 combined with Feature Pyramid Network (FPN) and achieved 95% accuracy for thorax, abdomen, and wings, with a mAP of 60% in validation and 52% in testing.

C. RQ3. What types of results are reported in studies on wing morphology?

Wing morphology has proven to be a valuable feature for automatic species classification. [22] developed a convolutional neural network (CNN) to identify seven species of Aedes from wing images. They used 1,155 images of Aedes and 554 non-Aedes mosquitoes, and trained CNNs in grayscale and RGB. This model achieved an F1-score of 99% for differentiating Aedes from other mosquitoes and around 90–91% for classifying the seven species, with 100% accuracy for Aedes albopictus. Classification errors occurred mainly among similar native species.

Comparing effectiveness, [23] evaluated full-body and wing images, finding that models trained with wing images achieved higher accuracy (87.6%) than with body images (78.9%) for the identification of four morphologically similar Aedes species. Wing-based models required fewer images for reliable performance. Likewise, model performance decreased significantly when evaluated with images from devices not included in training, although the study highlights the viability of body- and wing-based classification methods.

[24] evaluated the geometric morphometry of wings to identify six species of the genus Aedes in northeastern France. They used 18 reference points on the wings, applied Procrustes overlap analysis and Canonical Variant Analysis, achieving 98% accuracy in reclassification.

[25] developed an automatic system to identify Aedes species using wing interference patterns (WIPs) and deep learning. With a set of 494 images, they trained several CNN architectures, including MobileNet, ResNet18, and reduced versions of DarkNet. The models achieved 95% accuracy at the genus level, with perfect classification in half of the species.

[26] proposed a two-stage method for the automatic classification of midge species of the genus Culicoides based on morphological analysis of their wings. They applied image preprocessing techniques (filters and morphological operations) and machine learning, achieving 95.31% accuracy for wing segmentation and 94.79% for particle segmentation.

Additionally, [27] focused on the analysis of wingbeat patterns, presenting a hybrid method for the classification of mosquito species that combines different machine learning and deep learning architectures (SVM, MLP, Random Forest, Gradient Boosting, and kNN). They showed that hybrid architecture outperforms individual algorithms, as they achieved high accuracy and balanced performance in multi-class classification.

D. RQ4. What computational approaches have been proposed for morphological patterns in Diptera?

The integration of sophisticated architectures, such as Transformers, has set new standards for accuracy. [27] developed a deep learning model for automatic mosquito species identification based on Swin Transformer. They created a balanced dataset of 9,900 high-resolution images of 17 species and 3 subspecies. When comparing various convolutional networks and transformer-based models, the Swin Transformer-L variant was selected for its higher accuracy (called Swin MSI), which achieved 99.04% accuracy. In addition, this model demonstrated robustness by achieving 96.26% accuracy when classifying species not included in the training.

Along the same lines of advanced models, [29] developed a deep learning image analysis method to identify eleven mosquito species in Korea. They trained and compared five object detection models: Faster R-CNN with Swin Transformer, YOLOv5, ResNet101 DC-5, ResNeXt 101, and

RetinaNet. The results showed that the combination of Swin Transformer + Faster R-CNN achieved the highest accuracy with an F1-score of 97.1%, and YOLOv5 with 96.4%. They found that the combined use of RGB and fluorescent images, together with the non-maximum suppression (NMS) technique, improved the performance of all models. They identified the small sample size in some species as a limitation.

Finally, [29] developed a system to identify species using convolutional neural networks with a multilevel model that detects unknown species. They used a database of 12,977 images of 2,696 wild species, many with morphological damage. The system combines CNNs for feature extraction with classifiers (SVM, Random Forest) and a Gaussian mixture model for low-confidence cases. It achieves 97.04% accuracy in classifying 16 known species and 89.50% accuracy in detecting novel species.

One of the trends for future research is to improve the robustness and generalization of deep learning models. The study by [1] already highlighted that, despite using a large field dataset from the "Mosquito Alert" initiative, classification errors were directly related to poor image quality. This points to the need to develop models that are more robust to variability. Future research should focus on: preprocessing and data augmentation techniques; image quality detection models that can filter or warn about problematic images before classification and data collection, as suggested by [29] when identifying the small sample size in some species as a limitation, and the suggestion to improve the capture process to increase the volume of training data.

There is a trend to further investigate and validate the effectiveness of specific anatomical parts as descriptors, particularly wings. Findings from [22], [23] demonstrated that models trained with wing images achieved higher accuracy and required less data than full-body images. This will aid research into: the systematic comparison between the geometric morphometrics of [25], wing interference patterns (WIPs, Cannet et al., 2023), and wing-based CNNs to determine the most efficient technique; the application of anatomical extraction frameworks [20] to isolate and improve the image quality of wings and legs before classification; and the exploration of new motion-based descriptors, such as the flapping pattern analysis proposed by [26].

The implementation of architectures such as Transformer-based models to improve the identification of unclassified species. The Swin MSI model [27] has already demonstrated superior performance (99.04% accuracy) and generalization ability (96.26% on unseen species). Future research will focus on: exploring hybrid models that combine the high accuracy of Transformers with the ability to classify unknown or low-confidence species, as did the multilevel model of [30], which achieved 89.50% accuracy in detecting novel species; optimizing the combination of different image modalities, following the example of [28] with the use of RGB and fluorescent images to improve performance.

V. Conclusions

This research presents a mapping of the literature on the application of convolutional neural networks (CNNs) for classifying dipteran morphology. A descriptive analysis of the included studies was used, based on which the predominant architectures, the anatomical parts used, and the most frequently employed computational approaches were mapped, providing a current overview of the field.

According to the results found, the literature relies on pre-trained CNN architectures, such as VGG-16 and ResNet, as well as object detection models such as YoloV5, Faster R-CNN, and Mask R-CNN, which have demonstrated good performance in classification and morphological detection tasks. These architectures are mainly used through transfer learning, which allows for robust results even in scenarios with limited datasets. In this regard, the analysis shows evidence that wing morphology-based approaches consistently report better performance metrics, with accuracy and F1 scores exceeding 90% in most of the consulted articles. This confirms that the wings constitute a highly discriminating anatomical region, allowing for more precise and efficient models compared to using the whole body. In contrast, using whole-body images tends to employ more complex pipelines, based on higher detection or other strategies, to handle visual and structural variability.

Methodologically, the mapping shows a predominance of direct classification approaches, complemented by detection and segmentation, especially when it is necessary to locate specific structures or work with damaged specimens. Articles were identified that study hybrid models based on CNNs and Transforms, which represent a promising line of research for capturing complex spatial relationships, although their adoption is limited. The findings of this mapping are useful for researchers in computer vision and machine learning, as well as for entomologists, biologists, and digital taxonomy specialists, as it provides an overview of current trends in CNNs applied to image classification of dipteran morphology. This mapping contributes to a better understanding of the current state of applications in dipteran morphological classification, providing a solid foundation for future research in medical entomology.

References

[1] Adhane, G., Dehshibi, M. M., & Masip, D. (2021). A deep convolutional neural network for classification of Aedes albopictus mosquitoes. *IEEE Access*, 9, 72681–72690.

[2] Elhamod, M., et al. (2022). Hierarchy-guided neural network for species classification. *Methods in Ecology and Evolution*, 13(3), 642–652.

[3] Zhou, Z., et al. (2023). EchoAI: A deep-learning based model for classification of echinoderms in global oceans. *Frontiers in Marine Science*, 10, 1147690.

[4] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.

[5] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6), 84–90.

[6] Oliveira, M. B., et al. (2025). Classification of animal species via deep neural networks and species distribution modeling: A systematic review. *Artificial Intelligence Review*, 58, 230.

[7] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.

[8] Litjens, G., et al. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42, 60–88.

[9] Miao, Z., et al. (2019). Insights and approaches using deep learning to classify wildlife. *Scientific Reports*, 9, 8137.

[10] Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[11] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems* (Vol. 28, pp. 91–99).

[12] Khanam, R., & Hussain, M. (2024). What is YOLOv5: A deep look into the internal features of the popular object detector. arXiv:2407.20892.

[13] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 779–788).

[14] Xie, S., et al. (2017). Aggregated residual transformations for deep neural networks. In Proceedings of the *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 5987–5995).

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

[16] Lin, T. Y., et al. (2017). Focal loss for dense object detection. In Proceedings of the *IEEE International Conference on Computer Vision (ICCV)* (pp. 2980–2988).

[17] He, K., et al. (2017). Mask R-CNN. In Proceedings of the *IEEE International Conference on Computer Vision (ICCV)* (pp. 2961–2969).

[18] Liu, Z., et al. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the *IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 10012–10022).

[19] Howard, A. G., et al. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861.

[20] Minakshi, M., et al. (2020). A framework based on deep neural networks to extract anatomy of mosquitoes from images. arXiv:2007.14725.

[21] Sauer, F. G., et al. (2024). A convolutional neural network to identify mosquito species (Diptera: Culicidae) of the genus Aedes by wing images. *Scientific Reports,* 14(1), 3094.

[22] Nolte, K., et al. (2025). Potentials and limitations in the application of convolutional neural networks for mosquito species identification using wing images. *PLoS Computational Biology*, 21(9), e1013435.

[23] Martinet, J.-P., et al. (2021). Wing morphometrics of Aedes mosquitoes from north-eastern France. *Insects*, 12(4), 341.

[24] Cannet, A., et al. (2023). Wing interferential patterns (WIPs) and machine learning for the classification of some Aedes species of medical interest. *Scientific Reports*, 13(1), 11956.

[25] Venegas, P., et al. (2020). An approach to automatic classification of Culicoides species by learning the wing morphology. *PLoS ONE*, 15(11), e0241798.

[26] Gireesh, A., & Noortaj, S. (2025). Hybrid machine learning approach for mosquito species classification using wing beat analysis. International Journal of Scientific Research in Science, Engineering and Technology, 12(3), 1062–1070. https://ijsrset.com/index.php/home/article/view/IJSRSET2512124

[27] Zhao, D., et al. (2022). A Swin transformer-based model for mosquito species identification. *Scientific Reports*, 12(1), 18664.

[28] Lee, S., Kim, H., & Cho, B.-K. (2023). Deep learning-based image classification for major mosquito species inhabiting Korea. *Insects*, 14(6), 526.

[29] Goodwin, A., et al. (2021). Mosquito species identification using convolutional neural networks with a multitiered ensemble model for novel species detection. *Scientific Reports*, 11(1), 24040.

AUTHORS

Benjamín Mendoza

Benjamín Paulino Mendoza Contreras es estudiante de octavo semestre de la Licenciatura en Estadística en la Universidad Veracruzana, con sede en Xalapa, Veracruz, México. Sus intereses de investigación se centran en el aprendizaje automático (machine learning) y el aprendizaje profundo (deep learning), particularmente en el desarrollo y aplicación de modelos estadísticos y computacionales para el análisis de datos. Actualmente participa en actividades académicas relacionadas con la ciencia de datos y la modelación estadística.

Emmanuel Morales

Licenciado en Ciencias y Técnicas Estadísticas y Especialista en Métodos Estadísticos por la Universidad Veracruzana, con Maestría en Ciencias de la Información Geoespacial por el Centro Geo, CDMX (Centro CONAHCYT), actualmente cursando el Doctorado en Ciencias de la Computación en la Universidad Veracruzana. Profesor en la Licenciatura en Estadística, en la Especialidad en Métodos Estadísticos y la Maestría en Economía y Sociedad de China y América Latina en la misma universidad, donde he dirigido 15 tesis de licenciatura y 4 de especialidad. También tengo experiencia como Analista Estadístico en la Oficina del Programa de Gobierno del Estado de Veracruz. Mis líneas de investigación incluyen metodologías de cómputo, programación estadística, estadística multivariada, análisis espacial, ciencia de datos y modelos estadísticos, con aplicaciones en biología, medicina, ciencias administrativas y sociales. He participado en diversos congresos nacionales e internacionales.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026

AUTHORS

Cecilia Cruz

Profesora de tiempo completo en la Facultad de Estadística e Informática de la Universidad Veracruzana (UV). Es Doctora en Investigación Educativa por la UV, Maestra en Ciencias con especialidad en Estadística Aplicada por el ITESM Campus Monterrey, así como Especialista en Métodos Estadísticos y Licenciada en Estadística por la UV. Actualmente coordina la Especialización en Métodos Estadísticos y pertenece al Sistema Nacional de Investigadores (SNII) como Candidata.

Sus líneas de investigación abarcan la Educación Estadística, la Metodología Estadística Aplicada y la integración de la estadística con tecnologías emergentes como Machine Learning, Ciencia de Datos y Análisis Espacial. Ha colaborado en proyectos sobre sustentabilidad, alfabetización digital y actitudes hacia la estadística en estudiantes latinoamericanos.

Entre sus publicaciones recientes destacan trabajos sobre aprendizaje supervisado, formación en consultoría estadística y aplicaciones multivariantes, además de capítulos sobre alfabetización digital y redes sociales. Ha dirigido más de 70 tesis, combinando docencia, investigación e impulso al uso estratégico de la estadística para el desarrollo sostenible.

Luis Gomez

Es doctor en Administración y doctor en Finanzas Públicas. Estudió la Maestría en Impuestos, la Maestría en Contabilidad, así como la Especialidad en Administración. Es licenciado en Contaduría Pública Certificado por el IMCP y Licenciado en Psicología.

En su experiencia docente se desempeña como coordinador del Posgrado en Administración modalidad Virtual, asimismo es integrante del Cuerpo Académico Consolidado "Las organizaciones y su entorno". Investigador de Tiempo Completo del Instituto de Investigaciones y Estudios Superiores de las Ciencias Administrativas. El Dr. Luis Enrique es miembro del sistema nacional de investigadores e investigadoras, miembro certificado por Conocer, perfil deseable PRODEP y nivel 6 de productividad UV.

Además de ser autor y coautor de varios libros y revistas.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026

PBI-BFS-MaOA: A Many-Objective Evolutionary Algorithm with PBI-Based Boundary-Front Selection

ARTICLE HISTORY
Received 8 March 2026
Accepted 26 May 2026
Published 7 July 2026

Thiago Santos
Federal University of Ouro Preto (UFOP)
Associate Professor
Ph.D. in Mathematics
Brazil
santostf@ufop.edu.br
ORCID: 0000-0002-2435-2786

Sebastião Xavier
Federal University of Ouro Preto (UFOP)
Associate Professor
Ph.D. in Mathematics
Brazil
semarx@ufop.edu.br
ORCID: 0009-0004-2765-0764

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026

PBI-BFS-MaOA: A Many-Objective Evolutionary Algorithm with PBI-Based Boundary-Front Selection

Thiago Santos
Federal University of Ouro Preto (UFOP)
Associate Professor
Ph.D. in Mathematics
Brazil
santostf@ufop.edu.br

Sebastião Xavier
Federal University of Ouro Preto (UFOP)
Associate Professor
Ph.D. in Mathematics
Brazil
semarx@ufop.edu.br

***Abstract*—Reference-guided many-objective evolutionary algorithms often lose selection pressure when Pareto dominance becomes scarce and the final accepted front must be truncated. We propose PBI-BFS-MaOA, a many-objective evolutionary algorithm that preserves Pareto ranking for feasible solutions and modifies only the survival decision on the boundary front. The method combines cumulative ideal–nadir normalization, penalty-based boundary intersection association, active-niche filtering, and occupancy-aware survivor insertion. These operations are activated where convergence and directional coverage must be decided simultaneously. We evaluate the algorithm on DTLZ1–DTLZ4 and WFG1–WFG4 with $M \in \{5, 8, 10\}$ objectives, using the averaged Hausdorff distance Δ_p, Wilcoxon signed-rank tests, Friedman rank analysis, and runtime measurements. PBI-BFS-MaOA obtains the best mean Δ_p in 13 of 24 benchmark instances, with its strongest gains on high-dimensional DTLZ cases and degenerate WFG3 instances, while its runtime remains between NSGA-III and CMOEA-CD.**

***Keywords*—many-objective optimization, environmental selection, reference directions, penalty-based boundary intersection**

I. Introduction

Many-objective optimization is difficult not only because more criteria must be evaluated, but because Pareto dominance rapidly loses its ability to separate candidate solutions. Li et al. [1] identify this loss of discrimination as a central obstacle once the number of objectives moves beyond the classical two- or three-objective setting. Santos and Takahashi [2] give a formal account of the same phenomenon: as objective dimensionality increases, the probability that one candidate dominates another decreases sharply. The population then accumulates mutually non-dominated solutions, and environmental selection must choose among candidates that the initial dominance relation no longer orders with enough contrast.

This loss of contrast changes the role of survival selection. After the clearly superior fronts have been accepted, the remaining population slots are often disputed by an overflowing boundary front. At that point, the algorithm must preserve convergence while still maintaining directional coverage. The problem is geometric as much as statistical. In high-dimensional objective spaces, local density is harder to estimate, front shapes are harder to interpret, and unsupported reference directions may receive attention simply because the selection rule lacks a sharper local signal. Pal et al. [3] discuss a related difficulty from the perspective of objective reduction: deciding which information remains relevant becomes itself a nontrivial design problem.

Several algorithmic families address this pressure from different angles. Decomposition methods organize search through scalar subproblems, as in MOEA/D [4], and dominance-decomposition hybrids such as MOEA/DD [5] show that Pareto ordering and directional structure can coexist in the same survival mechanism. Reference-guided algorithms follow a related geometric logic. NSGA-III [6], θ-DEA [7], and RVEA [8] all use reference information to recover discrimination when non-dominated sets become too large. Their shared premise is clear: many-objective selection needs more than front rank.

Directional guidance, however, is not neutral. Ishibuchi et al. [9] show that decomposition-based performance is strongly affected by Pareto-front shape. A reference structure that works well on a regular front may become less reliable on disconnected, biased, or degenerate geometries. Qiu et al. [10], Liu et al. [11], Li et al. [12], and Wang et al. [13] respond to this issue by adapting reference structures or strengthening dominance with reference-vector information. These contributions suggest that survival quality depends not only on having directions, but on deciding which directions are actually supported by the current population.

The present work focuses on that decision at a narrower location: the boundary front. We do not replace Pareto ranking, redefine dominance globally, or introduce a multiarchive search architecture. Instead, PBI-BFS-MaOA preserves the usual Pareto scaffold for feasible solutions and intervenes only when the first overflowing front must be truncated. The proposed survival rule combines cumulative ideal–nadir normalization, penalty-based boundary intersection association, active-niche filtering, and occupancy-aware insertion. The aim is to use geometric information exactly where ordinary front ordering becomes underdetermined.

This local view has practical value in technological decision problems where many objectives must be balanced under a limited evaluation budget. Engineering design, energy dispatch, logistics planning, portfolio allocation, scheduling, and resource management often require simultaneous trade-

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.04

offs among cost, reliability, risk, environmental impact, and service quality. In such settings, wasting evaluations on poorly supported directions can delay the discovery of usable compromises. A boundary-front rule that protects the global Pareto order while improving the last survival decision is therefore relevant beyond benchmark optimization.

We evaluate the proposed intervention on DTLZ and WFG benchmark families with $M \in \{5, 8, 10\}$ objectives in the PymooLab environment [14]. Performance is measured with the averaged Hausdorff distance Δ_p, Wilcoxon signed-rank tests, Friedman rank analysis, and runtime measurements. NSGA-III and CMOEA-CD serve as structurally distinct baselines: the first is the canonical reference-guided Pareto method, and the second is a recent archive-cooperation approach. Section II places the proposal in the literature, Section III defines the survival rule, Section IV reports the numerical protocol and results, and Section V concludes the paper.

II. Related Work

The many-objective literature can be organized around a common design question: once Pareto dominance no longer separates most candidates, where should additional discrimination enter the evolutionary pipeline? Some methods modify the global search decomposition, others adapt the reference structure, and still others revise the dominance relation. The present work belongs to the environmental-selection line, but it is useful to position it against these alternatives [1].

Decomposition-based methods were among the earliest scalable responses. Zhang and Li [4] distributed the search across scalar subproblems, thereby reducing reliance on global pairwise dominance. Li et al. [5] later combined dominance and decomposition in MOEA/DD. This line of work established a key principle for many-objective search: front rank and directional scalarization can operate at different resolutions. Pareto order can maintain the broad convergence scaffold, while a directional mechanism can resolve local competition among candidates that are otherwise difficult to separate.

Reference-guided methods express the same principle in geometric form. Deb and Jain [6] used reference points to guide environmental selection in NSGA-III, which remains the most direct baseline for a modified reference-based survival stage. Yuan et al. [7] introduced angular information through θ-DEA, and Cheng et al. [8] built RVEA around reference-vector-guided survival. These algorithms differ in implementation, but they share a central conclusion: once non-dominated sets become too large, directional information is needed to make survival decisions operational.

That conclusion must be qualified. Ishibuchi et al. [9] show that decomposition-based algorithms are highly sensitive to Pareto-front shape. A fixed reference structure may be effective on regular fronts and less reliable on disconnected, biased, or degenerate fronts. Qiu et al. [10] improved objective-space decomposition under this concern, while Liu et al. [11], Li et al. [12], and Wang et al. [13] introduced self-guided, redistributed, or reference-vector-based dominance mechanisms. The common lesson is that the reference structure should not be treated as automatically valid everywhere in the objective space.

A second line preserves Pareto semantics while weakening or extending its strict form. Zhu et al. [15] generalized Pareto optimality for many-objective search. Tian et al. [16] strengthened dominance by combining convergence and diversity information, and Zhu et al. [17], [18] later developed generalized or relaxed dominance as a broader design framework. These methods show that dominance can be repaired, but they often act globally even though the strongest ambiguity may occur at a specific survival stage.

Environmental-selection studies make that local ambiguity explicit. Cheng et al. [19] argued that mating and environmental selection should be designed jointly, because the quality of selected parents is inseparable from the survival pressure they face. Sharma and Shukla [20] studied line-prioritized normalization and survivor choice, Myszkowski and Laszczyk [21] investigated diversity-based selection under constraints, and Liu et al. [22] treated environmental selection through clustering. These contributions differ in mechanics, but they agree on a practical point: the final accepted front is not a bookkeeping remainder. It is often where convergence and spread are either preserved or lost.

CMOEA-CD [23] provides a recent contrast. Instead of modifying one stage of selection, it uses three collaborative archives: a forward-exploration archive, a diversity-enhancement archive, and a feasibility-exploitation archive. This architecture separates exploration, diversity recovery, and feasible-solution intensification across different population-management channels. It is broader than the design studied here, and for that reason it is a useful comparator. Our question is more restricted: can a focused intervention in the boundary front recover selection pressure without replacing the surrounding evolutionary framework?

PBI-BFS-MaOA is therefore positioned between reference-guided selection and environmental-selection refinement. It keeps the Pareto-front order intact, associates only the critical front with normalized PBI directions, filters directions that lack support from the leading front when the front coverage is clearly sparse, and inserts survivors with an occupancy-aware rule. The contribution is intentionally local. Its value is not that it redesigns the entire many-objective algorithm, but that it targets the point at which the standard reference-guided survival rule becomes least decisive.

III. Proposed Method

The proposed method is a generational many-objective evolutionary algorithm whose main contribution lies in environmental selection rather than reproduction. Sampling, mating, and variation follow a classical evolutionary backbone. The restricted question is whether survival selection can recover discrimination inside the boundary front without discarding the global Pareto scaffold. To do so, the method retains front-based ordering where the population already provides a clear rank structure and intervenes only when the first overflowing front must be trimmed. The implementation then applies cumulative ideal–nadir normalization, PBI-based association, active-niche filtering, and occupancy-aware insertion. The method is therefore a survival-stage refinement within the standard evolutionary loop, focused on the point where dominance contrast is weakest in practice.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.04

Algorithm 1 Framework of the implemented PBI-BFS-MaOA

Input: population size N, reference directions $\mathcal{W}$, penalty parameter θ
Sample and evaluate the initial population $\mathcal{S}_0$
Initialize $\mathbf{z}^{\min}$ and $\mathbf{z}^{\max}$ from $\mathcal{S}_0$
Apply environmental selection to obtain $\mathcal{S}_0$ and its tournament fitness
while the stopping criterion is not met **do**
  Select parents by tournament using constraint violation and current fitness
  Generate offspring $\mathcal{Q}_t$ with a standard genetic variation operator
  Form the merged population $\mathcal{P}_t = \mathcal{S}_t \cup \mathcal{Q}_t$
  Update $\mathbf{z}^{\min}$ and $\mathbf{z}^{\max}$ from $\mathcal{P}_t$
  Apply environmental selection on $\mathcal{P}_t$
  Obtain the next population $\mathcal{S}_{t+1}$ and the updated tournament fitness
end while
Extract the current approximation set by filtering the population for feasible nondominated solutions

A. Framework of the Implemented Method

Let $\mathcal{S}_t$ denote the population at generation t, with $|\mathcal{S}_t| = N$, and let $\mathcal{W} = \{\mathbf{w}_1, \ldots, \mathbf{w}_K\}$ be the externally supplied set of reference directions. The process begins by sampling and evaluating an initial population, from which cumulative ideal and nadir estimates are initialized. Each generation then performs tournament-based parent selection, offspring generation by a standard variation operator, parent–offspring merging, and one environmental-selection call that returns both the survivors and the rank-based fitness values reused in the next tournament selection. Algorithm 1 summarizes the implemented architecture, and Fig. 1 presents the same process from the viewpoint of the survival decision.

The boundary-front rule is active from the first environmental-selection call onward, rather than being introduced as a late-stage correction. This matters because many-objective runs may produce large non-dominated subsets well before the final generations. The method is therefore directed at the $M > 3$ regime, where the last accepted front can dominate the survival outcome. The selection path moves through feasibility filtering, Pareto-front insertion, critical-front association, active-niche checking, and occupancy-aware insertion.

B. PBI-Based Boundary-Front Selection

Given the merged population $\mathcal{P}_t$, aggregated constraint violation is computed for each candidate $x_i \in \mathcal{P}_t$ as

$$\mathrm{CV}(x_i) = \sum_j \max\{0, g_j(x_i)\}. \tag{1}$$

This value induces the feasible subset

$$\mathcal{P}_t^f = \{x_i \in \mathcal{P}_t : \mathrm{CV}(x_i) = 0\}. \tag{2}$$

If $|\mathcal{P}_t^f| < N$, all feasible solutions are retained and the remaining slots are filled by the least infeasible candidates ranked by increasing CV. This branch is part of the implemented survival routine, although the numerical analysis in this paper uses unconstrained benchmark families. Its role is to establish a clear priority order: feasibility is handled first when feasible points are scarce, and the directional boundary-front rule is invoked only when enough feasible candidates exist for truncation to become the dominant issue.

When $|\mathcal{P}_t^f| \geq N$, selection proceeds on the feasible subset. The algorithm performs non-dominated sorting on the original objective vectors of $\mathcal{P}_t^f$ and obtains ordered fronts $\mathcal{F}_1, \mathcal{F}_2, \ldots$. Complete fronts are copied into the next population until the first overflowing front $\mathcal{F}_\ell$ is reached. Thus, the global convergence scaffold remains Pareto-ordered. The directional rule does not reshuffle clearly superior fronts; it acts only where Pareto ranking no longer determines the remaining survivor set.

The candidates in the critical front are then evaluated in normalized objective space. Let $\mathbf{z}_t^{\min}$ and $\mathbf{z}_t^{\max}$ denote the cumulative ideal and nadir estimates updated across generations. For an objective vector $\mathbf{f}_i$, the normalized vector is

$$\tilde{\mathbf{f}}_i = \frac{\mathbf{f}_i - \mathbf{z}^{\min}}{\max(\mathbf{z}^{\max} - \mathbf{z}^{\min}, \epsilon)}, \tag{3}$$

where ϵ is a componentwise safeguard against null spans. Each reference direction is normalized to unit length, and for every candidate–direction pair the algorithm computes

$$d_1(i,k) = \tilde{\mathbf{f}}_i^\top \mathbf{w}_k, \qquad d_2(i,k) = \sqrt{\|\tilde{\mathbf{f}}_i\|_2^2 - d_1(i,k)^2}. \tag{4}$$

The corresponding penalty-based boundary-intersection score is

$$\mathrm{PBI}(i,k) = d_1(i,k) + \theta_M d_2(i,k), \tag{5}$$

where

$$\theta_M = \begin{cases} \theta, & M \leq 3, \\ \theta\sqrt{M/3}, & M > 3. \end{cases} \tag{6}$$

Cumulative ideal–nadir normalization keeps directional comparisons stable across generations instead of allowing transient objective spans to dominate the association step. The PBI score separates an axial component d_1 from an orthogonal component d_2, so each candidate is evaluated by both progress along a direction and deviation from that direction. The scaling in (6) increases the orthogonal penalty as M grows, which is consistent with the stronger ambiguity observed in higher-dimensional objective spaces. The associated niche of x_i is then

$$k^\star(i) = \arg\min_k \mathrm{PBI}(i,k). \tag{7}$$

This association does not treat each direction as an independent scalar subproblem in the MOEA/D sense [4]. It provides a local geometric view only for candidates that have already passed Pareto-front screening.

C. Active-Niche Filtering and Occupancy-Aware Insertion

Once the complete fronts have been accepted, let $\mathcal{A}_t$ denote survivors already inserted before truncating the boundary front $\mathcal{F}_\ell$. Current occupancy is calculated for every niche k:

$$c_k = |\{x_i \in \mathcal{A}_t : k^\star(i) = k\}|. \tag{8}$$

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.04

LATIN-AMERICAN JOURNAL OF COMPUTING (LAJC), Vol XIII, Issue 2, July 2026

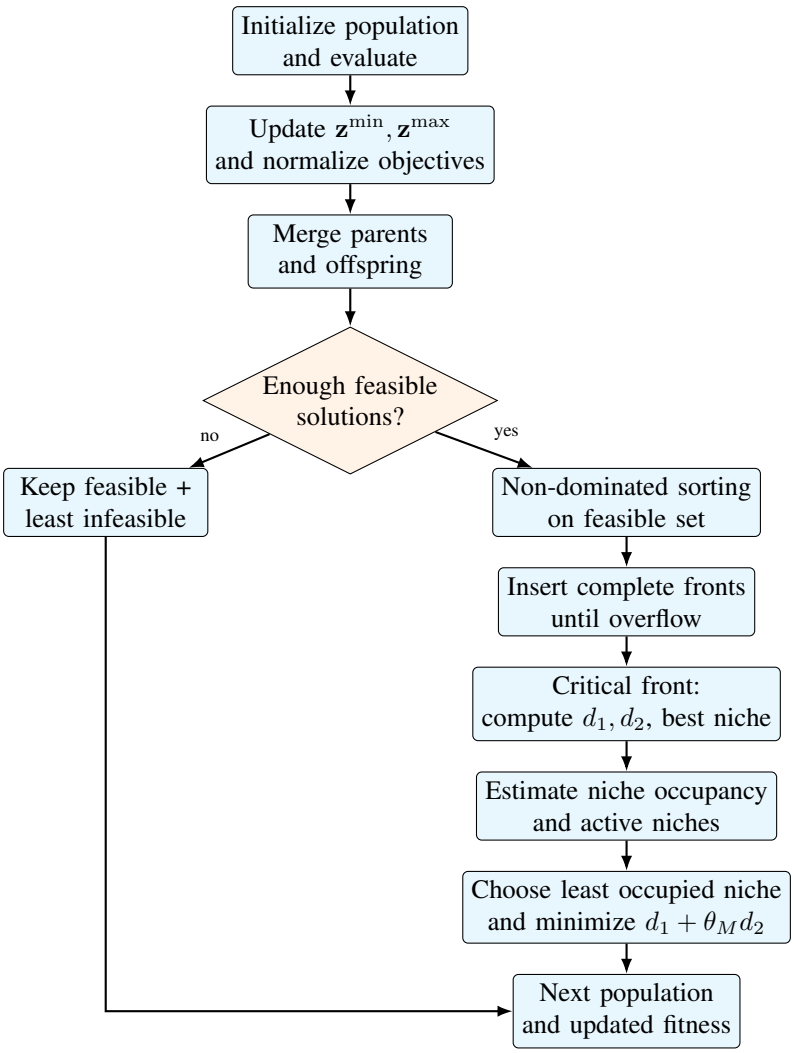


Fig. 1: Conceptual flow of the implemented PBI-BFS-MaOA survival mechanism

This occupancy count drives the boundary-front insertion rule. Additionally, the niche set that the first Pareto front activates is extracted in the algorithm:

$$\Omega_t = \{k^\star(i) : x_i \in \mathcal{F}_1\}. \tag{9}$$

If $|\Omega_t| < 0.8K$, the implementation treats the current leading front as insufficiently spread over the reference set. In that case, when the intersection is nonempty, only niches represented by $\mathcal{F}_\ell$ and active in Ω_t are admissible. Otherwise, all boundary-front niches remain admissible. The first front therefore acts as the support signal: if the best feasible solutions occupy only a restricted subset of directions, the algorithm avoids spending survivor slots on directions that the current population does not substantively support.

The $0.8K$ threshold is a conservative coverage gate rather than a tuned benchmark-specific parameter. It activates filtering only when at least 20% of the reference directions are unsupported by the leading front. This choice separates broadly covered fronts from fronts with visibly sparse directional support while avoiding the overly aggressive behavior that would arise from filtering whenever a small number of directions is missing. The value is also easy to interpret operationally: active-niche filtering is used only when the first front no longer represents most of the reference scaffold.

The last rule for insertion acts one survivor at a time. Every step identifies the least-occupied admissible niches, breaks ties randomly, and selects one niche. Let $\mathcal{C}_k \subseteq \mathcal{F}_\ell$ denote the candidates now assigned to the chosen niche k. The selected candidate is

$$x^\star = \arg\min_{x_i \in \mathcal{C}_k} \left(d_1(i, k^\star(i)) + \theta_M d_2(i, k^\star(i))\right). \tag{10}$$

This rule couples two pressures that in many-objective survival are often in tension. Poor niche occupancy allows for angular spread throughout the reference structure, and the within-niche PBI minimization favors candidates with better local convergence relative to the selected direction. The design is deliberately local. It does not change the ranking relation for the entire population, nor does it try to find a new global density model. It resolves the specific ambiguity created by the critical front after the better-ranked fronts have already been accepted.

After insertion, survivors receive rank-based fitness values according to their Pareto-front index. These values, together with constraint violation, are reused by tournament selection. Offspring generated by the variation operator are then merged back into the population, and the same survival rule is applied again. The empirical behavior reported in this paper should

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.04

therefore be read as the effect of a repeated boundary-front decision, not as a one-time tie-breaking operation.

The additional computational cost is concentrated in the critical-front association and insertion stage. Let $n_f = |\mathcal{P}_t^f|$ be the number of feasible candidates, $b = |\mathcal{F}_\ell|$ the size of the boundary front, K the number of reference directions, M the number of objectives, and r the number of remaining survivor slots. Non-dominated sorting over the feasible subset follows the usual front-ranking cost. The proposed modification adds $O(bKM)$ operations for normalized PBI association, $O(n_f + K)$ operations to compute current niche occupancy and active niches, and up to $O(r(b + K))$ operations for iterative occupancy-aware insertion under a direct implementation. Since $r \leq b$, the added survival-stage cost is bounded by $O(bKM + b^2 + bK + n_f)$. This is higher than the simplest NSGA-III niching pass, but it remains localized to the boundary front and is consistent with the observed runtime profile: slower than NSGA-III, yet substantially cheaper than the broader multiarchive management used by CMOEA-CD.

D. Design Highlights and Current Scope

The proposed approach has four active design choices. First, it keeps Pareto-front ordering as the global convergence scaffold and reserves directional scalarization for the moment at which that scaffold no longer determines the survivor set. Second, it uses cumulative ideal–nadir normalization to keep PBI association numerically stable across the run without excessive sensitivity to transient population spans. Third, it scales the orthogonal PBI penalty by $\sqrt{M/3}$ at $M > 3$, directly targeting the many-objective regime for which the method is designed. Fourth, it combines active-niche filtering and occupancy-aware insertion, making the boundary-front decision dependent on both local scalarization values and the current elite support of the reference structure.

These design choices also define the limits of the present claim. The constructor exposes the value σ_d, and the source file includes a helper for niche-penalty calculation, but in the current implementation that helper does not take part in the active survival path. The method is therefore not a fuzzy-dominance framework, a radial-repulsion mechanism, or a multiarchive strategy. The contribution, as evidenced in the code and by the benchmarks, is narrower and more specific: a Pareto-ordered many-objective evolutionary approach with normalized PBI-based boundary-front selection, conditional active-niche filtering, and occupancy-aware survivor insertion.

IV. Numerical Simulation and Analysis

A. Benchmark Test Problems

For the numerical study, we use two benchmark groups, namely DTLZ1–DTLZ4 [24] and WFG1–WFG4 [25], with objective counts $M \in \{5, 8, 10\}$. These settings place each experiment in the many-objective regime for which the proposed survival rule was developed. The benchmark families are deliberately complementary. DTLZ provides canonical and relatively interpretable front structures, while WFG introduces stronger distortions in shape, modality, deceptiveness, and degeneracy. Taken together, the two suites test whether the mechanism works only on regular fronts or also under geometries where reference-guided truncation is more difficult.

The DTLZ family isolates complementary sources of difficulty. DTLZ1 has a linear Pareto front with strong multimodality, making it useful for testing whether the boundary-front rule remains stable under competing local attractors. DTLZ2 has a smooth spherical front and provides a cleaner directional-coverage test. DTLZ3 retains the DTLZ2 geometry but adds severe multimodality, making it a difficult convergence test without changing the fundamental front shape. DTLZ4 biases the mapping toward extreme regions and is relevant for methods that depend on directional balance.

The WFG suite directs the evaluation toward more irregular geometries. WFG1 adopts bias and mixed shape transformations, while WFG2 generates disconnected and deceptive structures. WFG3 produces degenerate fronts, and WFG4 adds strong multimodality. WFG3 is particularly relevant because degeneracy is the scenario in which active-niche filtering should have its clearest effect.

We vary the number of objectives from 5 to 8 and 10 to test whether the local survival logic becomes more relevant as the selection stage grows more crowded.

B. Algorithms Comparison and Experimental Settings

The benchmark compares PBI-BFS-MaOA with two structurally relevant baselines: NSGA-III [6] and CMOEA-CD [23]. NSGA-III is the direct baseline because it combines Pareto fronts with reference directions and performs environmental selection through reference-guided niching. It is therefore the natural comparator for testing whether a new boundary-front rule improves a reference-guided survival protocol without changing the broader evolutionary paradigm.

CMOEA-CD is included for a different reason. Instead of relying on a single environmental-selection regime, it organizes a forward-exploration archive, a diversity-enhancement archive, and a feasibility-exploitation archive. Although CMOEA-CD was proposed for constrained multiobjective optimization, it remains informative here because it represents a recent and technically sophisticated alternative to single-stage survivor selection. The comparison therefore contrasts a local boundary-front intervention with a broader archive-cooperation design.

All algorithms are run in the PymooLab framework [14]. For every problem instance, the population size is fixed at 100, the maximum number of function evaluations is fixed at 50,000, and 30 independent runs are performed. The benchmark harness uses independent random seeds and records the indicator values and summary statistics used in the final performance evaluation.

NSGA-III and PBI-BFS-MaOA use the same reference-direction generation logic, which keeps the comparison focused on the survival operator rather than on different directional sets. For each instance, we report the mean and standard deviation of the performance indicator, the winning algorithm, the percentage gain of the winner over the closest comparator, and the pairwise Wilcoxon decision marker. This structure supports three readings: instance-wise comparison through the table, suite-wise comparison through Friedman

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.04

ranks, and computational-cost comparison through runtime statistics.

C. Experimental Results on Benchmark

The main performance indicator is Δ_p, the averaged Hausdorff distance between the approximation set returned by an algorithm and a reference Pareto front [26]. Let A denote the approximation set and let P^* denote the reference Pareto front. For a point u and a finite set S, define

$$d(u, S) = \min_{s \in S} \|u - s\|_2. \tag{11}$$

The p-averaged generational distance and inverted generational distance are written as

$$\mathrm{GD}_p(A, P^*) = \left(\frac{1}{|A|} \sum_{a \in A} d(a, P^*)^p \right)^{1/p}, \tag{12}$$

$$\mathrm{IGD}_p(A, P^*) = \left(\frac{1}{|P^*|} \sum_{p^* \in P^*} d(p^*, A)^p \right)^{1/p}. \tag{13}$$

Following Schütze et al. [26], the averaged Hausdorff distance is then

$$\Delta_p(A, P^*) = \max \left\{ \mathrm{GD}_p(A, P^*), \mathrm{IGD}_p(A, P^*) \right\}, \tag{14}$$

Lower Δ_p values indicate better performance because both convergence toward the reference front and spread along that front are penalized. The indicator is particularly relevant here because the proposed survival rule must negotiate convergence and diversity at the same stage.

A survival rule can reduce crowding while damaging local convergence, or improve convergence while collapsing directional coverage. Δ_p is sensitive to both failure modes and is therefore more stringent than a convergence-only indicator.

For each test instance, Table I reports the mean Δ_p value and sample standard deviation over 30 independent runs. Wilcoxon signed-rank tests are computed at significance level $\alpha = 0.05$ for pairwise comparisons, and Friedman tests are applied to compare average ranks across the DTLZ and WFG subsets.

The table is arranged as a single consolidated floating environment so that the full benchmark can be inspected at once. DTLZ and WFG are visually separated because the interpretation depends strongly on the distinction between canonical and distorted front geometries.

Table I shows that PBI-BFS-MaOA obtains the best mean Δ_p value in 13 of the 24 benchmark instances. NSGA-III is best in 7 cases, and CMOEA-CD is best in 4. The distribution of wins is more informative than the raw total. At $M = 5$, PBI-BFS-MaOA leads in only 2 of 8 cases, whereas at $M = 8$ and $M = 10$ it leads in $6/8$ and $5/8$ cases, respectively. This pattern agrees with the design motivation: as the number of objectives increases, the overflowing front becomes larger and less sharply separated under standard reference-guided selection, creating more room for a boundary-front intervention to help.

The DTLZ suite provides the clearest evidence in favor of the proposed approach. The Friedman test over the 12 DTLZ instances returns average ranks of 1.417 for PBI-BFS-MaOA, 1.833 for NSGA-III, and 2.750 for CMOEA-CD, with $\chi^2 = 11.167$ and $p = 3.76 \times 10^{-3}$. Pairwise Wilcoxon testing indicates a significant advantage of PBI-BFS-MaOA over CMOEA-CD ($p = 1.46 \times 10^{-3}$), while the difference from NSGA-III does not cross the 5% significance threshold ($p = 6.40 \times 10^{-2}$). Thus, on DTLZ, the proposed method clearly separates from CMOEA-CD and shows a repeated, although not statistically decisive, advantage over NSGA-III.

The strongest margins occur when multimodality and boundary-front pressure appear together. On DTLZ3 with $M = 10$, PBI-BFS-MaOA obtains $\Delta_p = 1.52$, compared with 21.9 for NSGA-III and 75.6 for CMOEA-CD. On DTLZ3 with $M = 8$, the corresponding values are 1.54, 14.3, and 19.8. For DTLZ1 with $M = 8$ and $M = 10$, the proposed method reaches 1.09×10^{-1} and 1.46×10^{-1}, compared with 2.00×10^{-1} and 4.71×10^{-1} for NSGA-III and much larger values for CMOEA-CD. These are precisely the cases in which many candidates can share the same front rank while differing substantially in directional plausibility.

The evidence is not uniform across all landscapes. NSGA-III remains marginally best on DTLZ1 with $M = 5$ and on DTLZ2 with $M = 5$ and $M = 10$, while CMOEA-CD obtains the best mean on DTLZ4 with $M = 5$. The WFG subset is also more balanced. The Friedman test does not reject comparable performance among the three algorithms on this subset ($\chi^2 = 2.000$, $p = 3.68 \times 10^{-1}$), and PBI-BFS-MaOA and NSGA-III have the same average rank, 1.833. This result is consistent with the fact that WFG includes disconnected, deceptive, degenerate, and highly multimodal structures, where different failure modes can dominate.

Even so, the WFG results support the specific role of active-niche filtering. PBI-BFS-MaOA obtains the best mean value on all three WFG1 instances and on WFG3 for $M = 8$ and $M = 10$. On WFG3 with $M = 10$, it obtains $\Delta_p = 7.23$, compared with 14.2 for NSGA-III and 15.2 for CMOEA-CD. On WFG3 with $M = 8$, the corresponding values are 2.60, 7.14, and 8.31. Since WFG3 is degenerate, these cases are aligned with the intended effect of active-niche filtering: when only a subset of directions is supported by the leading front, limiting insertion to active niches can prevent the algorithm from allocating survivors to weakly justified directions.

Runtime results support a balanced interpretation. Across instances, NSGA-III requires 10.69 s on average, PBI-BFS-MaOA requires 14.73 s, and CMOEA-CD requires 50.28 s. The proposed method is therefore approximately 37.8% slower than NSGA-III, but 70.7% faster than CMOEA-CD. Wilcoxon testing indicates that these runtime differences are systematic ($p < 1.2 \times 10^{-7}$ in each pairwise comparison). Across all 24 instances, the global Friedman test still favors PBI-BFS-MaOA, with average ranks of 1.625, 1.833, and 2.542 for PBI-BFS-MaOA, NSGA-III, and CMOEA-CD, respectively ($\chi^2 = 11.083$, $p = 3.92 \times 10^{-3}$).

Taken together, the results position PBI-BFS-MaOA as a quality–cost compromise. It is costlier than NSGA-III, substantially cheaper than CMOEA-CD, and frequently stronger in high-dimensional or degenerate settings where boundary-

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.04

TABLE I. Statistical results obtained by NSGA-III, CMOEA-CD, and PBI-BFS-MaOA on the DTLZ and WFG problems (Δ_p)

Problem	M	NSGA-III	CMOEA-CD	PBI-BFS-MaOA	Best	Gain (%)
DTLZ Suite						
DTLZ1	5	**6.63e-02 ± 2.8e-04**$^{=}$	2.73e-01 ± 3.9e-01^{-}	6.75e-02 ± 7.0e-04^{-}	NSGA-III	1.75
	8	2.00e-01 ± 1.7e-01^{-}	1.41e+00 ± 1.4e+00^{-}	**1.09e-01 ± 1.2e-03**$^{=}$	PBI-BFS-MaOA	45.19
	10	4.71e-01 ± 7.6e-01^{-}	4.44e+00 ± 3.7e+00^{-}	**1.46e-01 ± 1.2e-02**$^{=}$	PBI-BFS-MaOA	68.91
DTLZ2	5	**1.99e-01 ± 4.5e-05**$^{=}$	2.24e-01 ± 3.9e-03^{-}	1.99e-01 ± 4.7e-04^{-}	NSGA-III	0.17
	8	3.37e-01 ± 1.6e-04^{-}	4.92e-01 ± 2.0e-02^{-}	**3.35e-01 ± 7.4e-04**$^{=}$	PBI-BFS-MaOA	0.48
	10	**3.97e-01 ± 3.2e-04**$^{=}$	7.06e-01 ± 6.7e-02^{-}	4.03e-01 ± 1.8e-03^{-}	NSGA-III	1.32
DTLZ3	5	1.73e+00 ± 1.6e+00^{-}	1.52e+00 ± 1.5e+00^{-}	**3.79e-01 ± 5.4e-01**$^{=}$	PBI-BFS-MaOA	75.09
	8	1.43e+01 ± 7.3e+00^{-}	1.98e+01 ± 1.1e+01^{-}	**1.54e+00 ± 2.4e+00**$^{=}$	PBI-BFS-MaOA	89.21
	10	2.19e+01 ± 7.9e+00^{-}	7.56e+01 ± 6.6e+01^{-}	**1.52e+00 ± 1.5e+00**$^{=}$	PBI-BFS-MaOA	93.07
DTLZ4	5	2.42e-01 ± 1.1e-01$^{=}$	**2.17e-01 ± 3.8e-03**$^{=}$	2.61e-01 ± 1.2e-01$^{=}$	CMOEA-CD	16.80
	8	3.39e-01 ± 1.6e-02^{-}	4.51e-01 ± 9.7e-03^{-}	**3.36e-01 ± 8.2e-04**$^{=}$	PBI-BFS-MaOA	1.12
	10	4.07e-01 ± 2.3e-02$^{=}$	5.56e-01 ± 1.7e-02^{-}	**4.00e-01 ± 2.3e-03**$^{=}$	PBI-BFS-MaOA	1.78
WFG Suite						
WFG1	5	1.73e+00 ± 1.4e-02^{-}	1.94e+00 ± 1.4e-02^{-}	**1.69e+00 ± 1.8e-02**$^{=}$	PBI-BFS-MaOA	2.69
	8	2.47e+00 ± 6.3e-02^{-}	2.87e+00 ± 2.1e-02^{-}	**2.32e+00 ± 4.6e-02**$^{=}$	PBI-BFS-MaOA	6.01
	10	2.90e+00 ± 7.3e-02^{-}	3.36e+00 ± 4.0e-02^{-}	**2.65e+00 ± 7.7e-02**$^{=}$	PBI-BFS-MaOA	8.54
WFG2	5	**7.47e-01 ± 2.3e-01**$^{=}$	8.31e-01 ± 7.6e-02$^{=}$	8.64e-01 ± 2.9e-01$^{=}$	NSGA-III	13.56
	8	**1.93e+00 ± 6.5e-01**$^{=}$	2.24e+00 ± 9.6e-02^{-}	2.26e+00 ± 6.4e-01$^{=}$	NSGA-III	14.73
	10	**2.70e+00 ± 7.3e-01**$^{=}$	3.00e+00 ± 1.5e-01^{-}	2.73e+00 ± 8.3e-01$^{=}$	NSGA-III	1.09
WFG3	5	**1.67e+00 ± 2.4e-01**$^{=}$	2.67e+00 ± 3.8e-02^{-}	1.79e+00 ± 1.2e-01^{-}	NSGA-III	6.60
	8	7.14e+00 ± 8.5e-01^{-}	8.31e+00 ± 1.5e-01^{-}	**2.60e+00 ± 2.6e-01**$^{=}$	PBI-BFS-MaOA	63.53
	10	1.42e+01 ± 9.8e-01^{-}	1.52e+01 ± 1.8e-01^{-}	**7.23e+00 ± 5.4e-01**$^{=}$	PBI-BFS-MaOA	49.15
WFG4	5	1.72e+00 ± 8.7e-03^{-}	**1.50e+00 ± 3.9e-02**$^{=}$	1.74e+00 ± 6.8e-03^{-}	CMOEA-CD	13.65
	8	5.32e+00 ± 5.0e-02^{-}	**4.80e+00 ± 8.9e-02**$^{=}$	5.30e+00 ± 1.3e-02^{-}	CMOEA-CD	9.45
	10	7.73e+00 ± 1.2e-01^{-}	**7.46e+00 ± 1.1e-01**$^{=}$	7.54e+00 ± 3.2e-02^{-}	CMOEA-CD	1.14

front competition is prominent.

V. Conclusion and Future Work

This paper addressed many-objective optimization from a deliberately local standpoint. Instead of replacing the global Pareto scaffold, PBI-BFS-MaOA strengthens the decision made inside the final overflowing front. The method preserves Pareto ranking at the population level and introduces cumulative ideal–nadir normalization, PBI-based association, active-niche restriction, and occupancy-aware insertion only where standard survivor selection becomes least discriminative. Its contribution is therefore a boundary-front survival mechanism, not a new decomposition framework, a new generalized-dominance relation, or a multiarchive architecture.

The experimental evidence supports this design as a competitive alternative. Across the 24 benchmark instances, PBI-BFS-MaOA obtains the best overall Friedman rank and performs especially well on difficult DTLZ cases and on WFG3 as the number of objectives increases from 5 to 8 and 10. The strongest results occur in the more crowded configurations, where boundary-front competition is expected to be most severe. The WFG results are more mixed: NSGA-III remains strongest on WFG2, CMOEA-CD retains an advantage on WFG4, and the Friedman test on the WFG subset does not indicate a statistically significant separation among the three methods. Runtime places PBI-BFS-MaOA between the two baselines, so its practical value lies in a better quality–cost compromise than CMOEA-CD, although it does not match the lower computational cost of NSGA-III.

These findings suggest a practical recommendation for high-dimensional optimization tasks in engineering design, logistics, scheduling, resource allocation, and related decision systems: when a reference-guided algorithm repeatedly faces large boundary fronts, a localized PBI-based truncation rule can improve survivor quality without requiring a full redesign of the evolutionary framework. The method should not be read as a universal replacement for established reference-guided or archive-based approaches. It is better understood as a robust survival alternative for crowded or degenerate many-objective fronts.

Future work should follow three directions. First, a dedicated ablation study should separate the individual effects of dimensional scaling, active-niche filtering, and occupancy-aware insertion. Second, the feasibility branch defined through constraint violation should be evaluated on constrained DTLZ/WFG variants and engineering design benchmarks, since the present numerical study is limited to unconstrained problems. Third, sensitivity tests around the active-niche coverage threshold should be carried out to determine whether the conservative $0.8K$ gate remains appropriate across broader front geometries and population sizes.

Acknowledgment

The authors would like to thank METISBR: A Brazilian research group dedicated to Multi-Objective and Many-Objective Optimization (MaOPs) (https://github.com/METISBR), for the valuable discussions and the entire team's support during the development of this paper.

References

[1] B. Li, J. Li, K. Tang, and X. Yao, "Many-objective evolutionary algorithms: A survey," *ACM Computing Surveys*, vol. 48, no. 1, pp. 1–35, 2015.

[2] T. Santos and R. H. C. Takahashi, "On the performance degradation of dominance-based evolutionary algorithms in many-objective optimization," *IEEE Transactions on Evolutionary Computation*, vol. 22, no. 1, pp. 19–31, 2018.

[3] M. Pal, S. Saha, and S. Bandyopadhyay, "Decor: Differential evolution using clustering based objective reduction for many-objective optimization," *Information Sciences*, 2018.

[4] Q. Zhang and H. Li, "Moea/d: A multiobjective evolutionary algorithm based on decomposition," *IEEE Transactions on Evolutionary Computation*, vol. 11, no. 6, pp. 712–731, 2007.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.04

[5] K. Li, K. Deb, Q. Zhang, and S. Kwong, "An evolutionary many-objective optimization algorithm based on dominance and decomposition," *IEEE Transactions on Evolutionary Computation*, vol. 19, no. 5, pp. 694–716, 2015.

[6] K. Deb and H. Jain, "An evolutionary many-objective optimization algorithm using reference-point-based nondominated sorting approach, part i: Solving problems with box constraints," *IEEE Transactions on Evolutionary Computation*, vol. 18, no. 4, pp. 577–601, 2014.

[7] Y. Yuan, H. Xu, B. Wang, and X. Yao, "A new dominance relation-based evolutionary algorithm for many-objective optimization," *IEEE Transactions on Evolutionary Computation*, vol. 20, no. 1, pp. 16–37, 2016.

[8] R. Cheng, Y. Jin, M. Olhofer, and B. Sendhoff, "A reference vector guided evolutionary algorithm for many-objective optimization," *IEEE Transactions on Evolutionary Computation*, vol. 20, no. 5, pp. 773–791, 2016.

[9] H. Ishibuchi, Y. Setoguchi, H. Masuda, and Y. Nojima, "Performance of decomposition-based many-objective algorithms strongly depends on pareto front shapes," *IEEE Transactions on Evolutionary Computation*, vol. 21, no. 2, pp. 169–190, 2017.

[10] W. Qiu, J. Zhu, G. Wu, M. Fan, and P. N. Suganthan, "Evolutionary many-objective algorithm based on fractional dominance relation and improved objective space decomposition strategy," *Swarm and Evolutionary Computation*, 2021.

[11] S. Liu, Q. Lin, K.-C. Wong, C. A. C. Coello, J. Li, Z. Ming, and J. Zhang, "A self-guided reference vector strategy for many-objective optimization," *IEEE Transactions on Cybernetics*, vol. 52, no. 2, pp. 1164–1178, 2022.

[12] W. Li, Y. Chen, Y. Dong, and Y. Huang, "A solution potential-based adaptation reference vector evolutionary algorithm for many-objective optimization," *Swarm and Evolutionary Computation*, vol. 85, p. 101451, 2024.

[13] S. Wang, H. Wang, Z. Wei, F. Wang, Q. Zhu, J. Zhao, and Z. Cui, "A pareto dominance relation based on reference vectors for evolutionary many-objective optimization," *Applied Soft Computing*, vol. 152, p. 111505, 2024.

[14] T. Santos and S. Xavier, "Pymoolab: An open-source visual analytics framework for multi-objective optimization using llm-based code generation and mcdm," 2026, preprint available at https://arxiv.org/abs/2603.01345.

[15] C. Zhu, L. Xu, and E. D. Goodman, "Generalization of pareto-optimality for many-objective evolutionary optimization," *IEEE Transactions on Evolutionary Computation*, vol. 20, no. 2, pp. 299–315, 2016.

[16] Y. Tian, R. Cheng, X. Zhang, Y. Su, and Y. Jin, "A strengthened dominance relation considering convergence and diversity for evolutionary many-objective optimization," *IEEE Transactions on Evolutionary Computation*, vol. 23, no. 2, pp. 331–345, 2019.

[17] S. Zhu, L. Xu, E. Goodman, K. Deb, and Z. Lu, "A general framework for enhancing relaxed pareto dominance methods in evolutionary many-objective optimization," *Memetic Computing*, vol. 14, pp. 289–308, 2022.

[18] S. Zhu, L. Zeng, and M. Cui, "Symmetrical generalized pareto dominance and adjusted reference vector cooperative evolutionary algorithm for many-objective optimization," *Symmetry*, vol. 16, no. 11, pp. 1–22, 2024.

[19] J. Cheng, G. G. Yen, and G. Zhang, "A many-objective evolutionary algorithm with enhanced mating and environmental selections," *IEEE Transactions on Evolutionary Computation*, vol. 19, no. 4, pp. 592–605, 2015.

[20] D. Sharma and P. K. Shukla, "Line-prioritized environmental selection and normalization scheme for many-objective optimization using reference-lines-based framework," *Swarm and Evolutionary Computation*, 2019.

[21] P. B. Myszkowski and M. Laszczyk, "Diversity based selection for many-objective evolutionary optimisation problems with constraints," *Information Sciences*, 2021.

[22] S. Liu, J. Zheng, Q. Lin, and K. C. Tan, "Evolutionary multi and many-objective optimization via clustering for environmental selection," *Information Sciences*, vol. 578, pp. 930–949, 2021.

[23] Z. Liu, F. Han, Q. Ling, H. Han, and J. Jiang, "Constraint-pareto dominance and diversity enhancement strategy-based evolutionary algorithm for solving constrained multiobjective optimization problems," *IEEE Transactions on Evolutionary Computation*, vol. 29, no. 6, pp. 2771–2784, 2025.

[24] K. Deb, L. Thiele, M. Laumanns, and E. Zitzler, "Scalable test problems for evolutionary multiobjective optimization," in *Evolutionary Multiobjective Optimization: Theoretical Advances and Applications*. London, U.K.: Springer, 2005, pp. 105–145.

[25] S. Huband, L. Barone, L. While, and P. Hingston, "A scalable multi-objective test problem toolkit," *Lecture Notes in Computer Science*, vol. 4193, pp. 280–295, 2006.

[26] O. Schütze, X. Esquivel, A. Lara, and C. A. C. Coello, "Using the averaged hausdorff distance as a performance measure in evolutionary multiobjective optimization," *IEEE Transactions on Evolutionary Computation*, vol. 16, no. 4, pp. 504–522, 2012.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.04

AUTHORS

Thiago Santos

Thiago Santos is an Associate Professor at the Federal University of Ouro Preto (UFOP), Ouro Preto, Brazil. He holds a Ph.D. in Mathematics and coordinates both the Mathematics Education Research Group (GEEMA) and the Applied Mathematics Group of the Department of Mathematics. His research spans multi-objective optimization, evolutionary computation, and mathematics education, with special attention to the interaction between rigorous mathematical modeling and computational methods. In optimization and computational intelligence, his work focuses on the design and analysis of metaheuristic algorithms able to handle several conflicting objectives simultaneously, with applications in engineering and applied sciences. He also develops research in mathematics education, addressing pedagogical innovation, curriculum organization, and the conceptual barriers faced by students in advanced mathematical reasoning. Through this integrated agenda, he contributes to the theoretical foundations of optimization methods while supporting more effective approaches to university-level mathematics teaching. He is a founding member of the METISBR research group on multi-objective and many-objective optimization.

Sebastião Xavier

Sebastiao Xavier is an Associate Professor at the Federal University of Ouro Preto (UFOP), Brazil. He received his B.S., M.Sc., and Ph.D. in Mathematics from the Federal University of Minas Gerais (UFMG), developing specialized expertise in dynamical systems and real foliations. His academic career includes extensive teaching experience from basic education to graduate-level mathematics, together with sustained participation in the institutional development of mathematics programs at UFOP. Through the Mathematics Education Research Group (GEEMA), he has contributed to the training of future educators and to discussions on mathematical formation. His current scientific work is centered on optimization, especially multiobjective optimization and evolutionary strategies. In this area, he is interested in connecting rigorous theoretical foundations with computational procedures that can support practical decision-making. His research profile combines pure mathematics, applied optimization, educational engagement, and academic service. He is a member of the METISBR research group on multi-objective and many-objective optimization.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026

Mobile Applications in Mental Health and Public Safety: Challenges and Gaps in Digital Transformation

ARTICLE HISTORY
Received 24 December 2025
Accepted 27 March 2026
Published 7 July 2026

Diego Mattera
UCASAL
Buenos Aires, Argentina
mattera9@gmail.com
ORCID: 0009-0001-2937-7229

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026

D. Mattera, "Mobile Applications in Mental Health and Public Safety: Challenges and Gaps in Digital Transformation", Latin-American Journal of Computing (LAJC), vol. 13, no. 2, 2026.

Mobile Applications in Mental Health and Public Safety: Challenges and Gaps in Digital Transformation

Aplicaciones Móviles en Salud Mental y Seguridad Ciudadana: Desafíos y Brechas en la Transformación Digital

Diego Mattera
UCASAL
Buenos Aires, Argentina
mattera9@gmail.com

***Abstract*—**Mobile applications in mental health and public safety have evolved from individual tools into strategic digital infrastructures within the process of digital transformation. This study presents a narrative and documentary review focused on the United States context. The results show that, although these applications demonstrate high functional effectiveness and sustained growth, their social impact is limited by inequalities in access, digital literacy, institutional trust, and interoperability with public services. In mental health, app use among individuals with diagnosed disorders remains low, with moderate effects on reducing anxiety and depression symptoms and intermittent adherence, despite high levels of smartphone ownership. In public safety, platforms such as Life360 and Citizen report user growth; however, evidence regarding their real operational impact, response time reduction, and perceived security is limited and heterogeneous. The social sustainability of these technologies requires institutional validation, inclusive design, data protection, community participation, and alignment with public policies. The convergence with artificial intelligence, wearable devices, and hybrid intervention models projects future trends toward intelligent personalization and the strengthening of community resilience.

***Resúmen*—**Las aplicaciones móviles en salud mental y seguridad ciudadana han evolucionado de herramientas individuales a infraestructuras digitales estratégicas en la transformación digital. Este estudio realiza una revisión narrativa y documental, centrada en el contexto estadounidense. Los resultados muestran que, aunque estas aplicaciones presentan alta eficacia funcional y crecimiento sostenido, su impacto social se ve limitado por desigualdades en acceso, alfabetización digital, confianza institucional e interoperabilidad con servicios públicos. En salud mental, el uso de aplicaciones entre personas con trastornos diagnosticados es reducido, presenta efectos moderados en la reducción de síntomas de ansiedad y depresión y adherencia intermitente, pese a la alta propiedad de smartphones. En seguridad ciudadana, plataformas como Life360 y Citizen registran crecimiento de usuarios, pero la evidencia sobre impacto operativo real, reducción de tiempos de respuesta y percepción de seguridad es limitada y heterogénea. La sostenibilidad social de estas tecnologías requiere validación institucional, diseño inclusivo, protección de datos, participación comunitaria y articulación con políticas públicas. La confluencia con inteligencia artificial, dispositivos vulnerables y modelos híbridos de intervención proyecta tendencias futuras hacia personalización inteligente y fortalecimiento de la resiliencia comunitaria.

Keywords*—mobile applications, mental health, citizen security, digital divide, technology adoption, digital transformation.*

***Palabras clave*—** *aplicaciones móviles, salud mental, seguridad ciudadana, brecha digital, adopción tecnológica, transformación digital.*

INTRODUCTION

In recent decades, mobile applications have evolved from tools designed for individual convenience into critical infrastructures for managing public safety and mental health, within the framework of an accelerated process of digital transformation (Ventola, 2014). These technologies driven by the integration of artificial intelligence, geolocation, process automation, and community engagement have reshaped approaches to prevention, monitoring, and intervention in situations of individual and collective risk, expanding both institutional response capacities and digitally mediated forms of social participation.

In the field of mental health, several studies conclude that mobile apps enhance service coverage through digital therapies, emotional monitoring, and asynchronous interventions, particularly in contexts characterized by high demand and limited availability of in-person services (Luxton et al., 2016; Hwang et al., 2021; Havard et al., 2015). In parallel, platforms oriented toward public safety have demonstrated potential to strengthen real-time incident reporting, community coordination, and integration with emergency systems, fostering new dynamics of collective intelligence and distributed surveillance.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.05

Nevertheless, the expansion of these mobile applications occurs within a landscape marked by persistent structural barriers. Current literature highlights that the digital divide, technological fragmentation, limited interoperability with public systems, and risks related to sensitive data protection and algorithmic bias significantly constrain their effectiveness as instruments of social protection (Kitchin, 2014; WHO, 2022). Specifically, while research such as that of McCloud et al. (2020), published in JMIR Mental Health, reports significant reductions in symptoms of depression and anxiety through sustained use of digital therapy applications, notable inequalities have also been observed in terms of access, continuous use, and quality of outcomes. These disparities are largely determined by digital literacy, socioeconomic status, and institutional context. This tension between high functional potential and unequal structural conditions raises critical questions regarding the actual scope of mobile applications as sustainable instruments of social intervention.

Against this backdrop, the present article aims to analyze how mobile applications oriented toward mental health and public safety are being developed and implemented as tools for prevention, monitoring, and response to risk situations, as well as to examine the main technical, ethical, and institutional challenges that affect their effective integration with public systems. Although the scope of analysis is general, the empirical evidence focuses particularly on the case of the United States, given its high adoption of technological tools, availability of institutional data, and advanced development of digital infrastructures.

Through a narrative and documentary review of scientific literature, institutional reports, and representative digital platforms from both fields of study, this work seeks to contribute a strategic and critical perspective on the role of mobile applications as instruments of digital change, risk prevention, and collaborative governance, with particular attention to the structural inequalities that shape their adoption, impact, and social assimilation.

THEORETICAL FRAMEWORK: CONCEPTUAL, FUNCTIONAL, AND TECHNOLOGICAL APPROACHES

A. Mobile Applications and Digital Transformation

Recent academic production converges in conceptualizing mobile applications as central socio-technical devices in the current phase of digital transformation, as they structure technological developments, automation processes, and advances in mediation between individuals, institutions, and social protection systems.

In the domains of mental health and public safety, these mobile platforms not only fulfill instrumental functions but also operate as digital infrastructures for intervention, prevention, and risk management. From a systemic perspective, Luxton et al. (2016) argue that mental health applications reconfigure traditional care models by enabling asynchronous interventions, continuous monitoring of emotional states, and rapid access to therapeutic resources, thereby reducing limitations related to geographic location, professional availability, and access costs. In a similar vein, Firth et al. (2017) situate them within the field of digital mental health interventions, emphasizing their potential as technological extensions of traditional clinical systems in contexts of high demand.

B. Mental Health Applications

From a functional standpoint, mobile mental health applications can be grouped into three major categories:

Therapeutic support apps, which provide mood tracking, psychoeducation, and structured tactics for digital psychological intervention (e.g., Youper, Sanvello).

Emotional well-being and prevention apps, which focus on reducing stress, regulating emotional health, and improving sleep (e.g., Calm, Headspace, Insight Timer).

Crisis management apps, aimed at suicide prevention and urgent support, prioritizing immediate access to support networks and emergency services (e.g., MY3, NotOK).

This classification underscores the progressive expansion of the digital mental health field toward devices of prevention, intervention, and continuous support.

Building on this functional basis, mobile mental health applications are developed through diverse patterns of digital therapeutic mediation. On one hand, tele-psychotherapy platforms (e.g., Talkspace, BetterHelp) rely on multichannel communication schemes text, audio, and video that sustain technology-mediated therapeutic processes under conditions of greater temporal flexibility and spatial delocalization. On the other hand, these technologies also incorporate artificial intelligence (e.g., Woebot, Wysa), which employ conversational systems grounded in cognitive-behavioral therapy principles, oriented toward emotional self-regulation, psychological training, and early symptom detection. These advances fall within the domain of digital therapies, characterized by the fusion of clinical knowledge, automated processes, and the potential for large-scale expansion.

In this same line, the systematic review by Almuqrin et al. (2025) highlights sustained growth in evidence-based mental health applications. Nevertheless, significant barriers remain in their clinical validation, regulation, and effective translation into healthcare systems, requiring standardized criteria for their responsible incorporation.

C. Public Safety Applications

Recent literature on public safety highlights a progressive shift from patterns of targeted surveillance toward distributed surveillance schemes and citizen participation mediated by digital platforms.

From a functional perspective, public safety applications exhibit a clear differentiation:

Alert and reporting apps, which enable direct contact with emergency services and the georeferenced transmission of real-time information (e.g., official emergency applications).

Preventive personal safety apps focus on location monitoring, automatic alert activation in risk situations, and communication with trusted contacts (e.g., Life360, bSafe).

Collaborative surveillance apps, based on citizen participation for incident reporting, risk mapping, and information exchange among neighbors (e.g., Citizen, local neighborhood alert systems).

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.05

These typologies reflect the transition from reactive models to preventive schemes of digital risk management.

Several studies have noted that these technologies not only restructure urban self-protection practices but also reshape the social production of surveillance and territorial control. In this regard, Kitchin (2014) argues that the digital infrastructures of so-called "smart cities" reorganize traditional forms of security governance by implementing decentralized citizen participation, real-time decision-making processes, and algorithmic systems.

Consequently, these mobile applications are analyzed as risk communication infrastructures that enable decentralized information production, real-time incident reporting, coordination of family and community trajectories, and the generation of situational alerts (e.g., Citizen, Nextdoor, AlertCops). These platforms do not replace state security mechanisms but act as complementary layers of information, thereby expanding event recording capacity, accelerating response circuits, and reshaping social perceptions of protection. In this sense, security ceases to depend exclusively on vertical structures and increasingly incorporates horizontal dynamics of digital cooperation.

D. Technological Architecture and Functional Impacts

From a technical-structural standpoint, mobile applications in public safety and mental health share a modular, scalable, and user-centered architecture, sustained by the integration of multiple technological layers.

Key structural components include:

1. Artificial intelligence, oriented toward predictive analysis of emotional states, behavioral patterns, and risk scenarios.
2. Geolocation, supporting the identification of critical areas, spatial coordination of responses, and contextualized activation of alerts.
3. Process automation, through conversational bots, emergency protocols, and instant notification systems.
4. Advanced information security mechanisms, including end-to-end encryption, biometric authentication, and access control, which are essential in environments managing sensitive data, as documented by Kitchin (2014) and Lupton (2015) in their analyses of digital infrastructures applied to health and safety.

From a functional perspective, specialized literature documents that mental health applications enable, unprecedented forms of longitudinal monitoring of emotional states through systematic records, psychological assessment scales, affect regulation exercises, and crisis intervention tools (Berrouiguet et al., 2016; Lipschitz et al., 2022; Lehtimäki et al., 2021).

More broadly, hybridization is observed between technological logic, social intervention, and institutional risk management, characteristic of new regimes of digital governance, where platforms, algorithms, and mobile devices articulate institutional practices and social dynamics of prevention and control (Lupton, 2015; Kitchin, 2014; Lyon, 2018).

E. Ethical and Regulatory Tensions

Specialized literature warns that the expansion of these digital infrastructures is marked by complex ethical, regulatory, and political tensions. Among the main challenges are the vulnerability of personal data privacy, the opacity of algorithmic systems, biases in classification and prediction mechanisms, and technological fragmentation that limits interoperability between private platforms and public services.

Regarding algorithmic ethics, authors such as Mittelstadt et al. (2016) caution that automated decision-making systems may reproduce structural inequalities, generate unintended discriminatory effects, and complicate the attribution of responsibility—particularly when they operate as opaque "black boxes" for users and institutions.

These issues must be addressed in alignment with regulatory frameworks and international standards, such as the General Data Protection Regulation (GDPR), the Health Insurance Portability and Accountability Act (HIPAA) in the healthcare domain, and the recommendations of the National Institute of Standards and Technology (NIST) in cybersecurity. These frameworks aim to ensure principles of transparency, fairness, security, and algorithmic accountability in the deployment of such technologies.

METHODOLOGY

A. Study Design

This study adopts a narrative and documentary review approach, aimed at analyzing the adoption, impact, and challenges of mobile applications in the domains of mental health and public safety. This type of research design allows for the integration of scientific literature, sectoral reports, institutional documents, and data from representative digital platforms, providing a critical and comprehensive perspective on models of implementation and use of these technological tools. The choice of this approach is justified by the heterogeneity of available data, the diversity of application contexts, and the need to identify both general trends and structural constraints.

B. Source Selection and Inclusion Criteria

This work incorporates relevant sources published between 2015 and 2025, prioritizing studies with official reports, empirical evidence, and international documentation on digital health and emergency systems. The following were considered:

Scientific articles on mobile applications for mental health and public safety. Institutional reports from national and international organizations, including EENA, GAO, and U.S. government agencies. Data from representative technological platforms, such as Life360 and Citizen, to contextualize adoption and market penetration. Market studies and sectoral reports on expansion, demand, and functional coverage of apps.

Anecdotal sources, studies focused exclusively on technical aspects without social or functional implications, and reports lacking clear references were excluded.

C. Analytical Strategy

The analysis was structured around three main axes:

Adoption patterns and structural gaps: evaluation of inequalities in access and use, considering socioeconomic, geographic, cultural, digital literacy, and institutional trust factors.

Impact on mental health: review of evidence on the use of applications for psychological support, digital therapies, and

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.05

emotional monitoring, including prevalence of use, adherence, continuity, and reported outcomes.

Scope in public safety: analysis of the adoption of community surveillance and reporting apps, their relationship with official emergency systems (Next Generation 911), and a critical evaluation of their actual impact on response times and perceptions of safety.
A comparative and critical approach was applied, triangulating institutional reports and market data with academic studies, with particular attention to the United States as the central context.

D. Synthesis Procedure

Extracted information was organized into thematic matrices that enabled comparison across adoption patterns, functional impacts, and structural limitations.
The narrative synthesis combined quantitative results—such as market estimates, adoption percentages, and user coverage—with qualitative analysis of structural inequalities, digital exclusion factors, technical challenges, and ethical barriers.
Through this procedure, general trends and gaps in empirical evidence were identified, guiding the discussion of findings toward digital inclusion policies, participatory design approaches, and strategies for institutional strengthening.

E. Implementation of the Analysis

The narrative review was organized through a systematic process of searching, selecting, and synthesizing scientific literature, institutional reports, and documentation from relevant digital platforms. The collected data were coded and analyzed narratively, with emphasis on the functional potential and the structural barriers that condition the adoption, effectiveness, and social appropriation of mobile applications.
This approach enabled a comparative analysis across different contexts and technologies, as well as the identification of ethical, technological, and institutional challenges, together with opportunities to strengthen digital inclusion, interoperability, and collaborative governance.
The results of the review are presented below, organized according to the axes of adoption, impact on mental health, and scope in public safety. Socioeconomic inequalities directly affect processes of digital exclusion, as evidenced by the digital divide based on access to connectivity and technological devices, particularly among populations residing in rural areas or in contexts with precarious infrastructure, lower income, and lower educational attainment. This pattern has also been observed in the United States, where data reflect the impact of structural inequalities on low-income populations, rural residents, and racial minorities, who face greater difficulties in accessing available digital applications. Such asymmetry limits the capacity of health and safety applications to function as universal tools (Hernández & Roberts, 2018). The mere incorporation of digital technologies into the population does not guarantee homogeneous results in terms of use, especially in territories with limited infrastructure. In this respect, a recent study by Wen and Tian (2024) indicates that the mental health benefits associated with digital access are more evident in urban than in rural contexts.

Table I. Limiting Factors in Access to Mobile Health and Public Safety Applications

Dimension	Limiting Factor	Impact Description
Socio-educational	Low digital literacy	Hinders the effective use of mobile health, including mental health and public safety applications, particularly among populations with lower educational attainment.
Structural	Access inequalities	Limitations in access to mobile devices, connectivity, and digital services, which reduce the possibilities of technological adoption.
Sociocultural	Cultural and/or linguistic barriers	Linguistic and cultural differences hinder the understanding, appropriation, and proper use of mobile applications.
Technological-institutional	Distrust in data privacy	Fear of misuse of personal information discourages the download and use of digital platforms.

In the specific case of mobile applications oriented toward public safety, a study published in Police Practice and Research (Elphick et al., 2021) analyzed the privacy terms and conditions of 240 applications and revealed the existence of scarce—or even absent—regulations regarding data protection. This situation acts as a deterrent for users who distrust institutions. Most applications require registration or login; only slightly more than half (55%) of reporting apps allow anonymous submissions, and barely 10% provide comprehensible privacy policies. These characteristics generate vulnerability, particularly among individuals with lower digital competencies who face greater difficulties in accessing these tools equitably.

In the United States, deficits in privacy terms and institutional distrust particularly affect migrant communities, African Americans, and individuals with histories of police surveillance. In addition, economic inequalities limit access to safety applications that require constant connectivity, mobile data consumption, and updated devices. In this context, so-called "digital public protection" tends, in practice, to exclude those with fewer social advantages. The absence of access to mobile technologies constitutes the first level of the digital divide and has a direct impact on security conditions, as noted by Mihale-Wilson et al. (2025). At the same time, the increase in the use of neighborhood surveillance and reporting applications in the United States has reignited debates on privacy, distrust of authorities, selective surveillance, and algorithmic bias. Moreover, the collection of sensitive data under low or nonexistent protection standards does not effectively guarantee rights such as anonymization nor foster genuine community participation, contributing to the rejection of these technologies by marginalized or vulnerable communities (Elphick et al., 2021). In summary, the systems and literature reviewed demonstrate that the development and expansion of public safety applications hold significant potential to strengthen citizen participation, reduce

bureaucratic barriers in reporting systems, and optimize security mechanisms. However, their deployment must be accompanied by policies aimed at improving digital literacy, infrastructure, data protection, and technological inclusion. Furthermore, it is essential to promote designs that actively address existing economic and structural inequalities. To overcome the unequal conditions that determine who can benefit from technological innovations and access to health, coordinated interventions are required across institutional, social, and political domains.

A. Impact in the Field of Mental Health

Digital mental health applications can serve as complementary or alternative tools for psychological support. This potential is linked to data from the MHA 2023 report, which indicates that 20.8% of the U.S. adult population—over 50 million people—experienced a mental disorder during that year, and a significant proportion expressed dissatisfaction with the treatments received. Several studies conducted in recent years show that the rapid increase in smartphone adoption has not corresponded to an equivalent growth in the use of mental health applications among individuals with diagnosed disorders.

According to Deressa Guracho et al. (2023), the combined prevalence of app use among people with mental disorders was approximately 23.3%. In a survey of the general population, 41% of participants reported having used a mental health application in the past 12 months (Fürtjes et al., 2024), while another study indicated that in the United States, 43% accessed mental well-being applications, though only 18% used them for clinical purposes (Vera Cruz, 2023). Available empirical evidence suggests that these applications have moderate effects in reducing symptoms of anxiety and depression (Bell et al., 2022; Vera Cruz, 2023; Fürtjes et al., 2024).

However, adherence and continuity of use are often limited, which restricts their long-term therapeutic impact. In parallel, the global market for mental health applications is projected to grow steadily, with estimates reaching USD 17.5 billion by 2030 (Grand View Research, 2025).

Table II below presents a synthesis of the use and impact of mental health applications, according to recent studies.

Source	Indicator	Results
MHA 2023 [22]	Prevalence of mental disorders in U.S. adults	20.8% of adults (50 million people)
Deressa Guracho et al. [23	Use of apps among individuals with diagnosed mental disorders	23%
Fürtjes et al. [24]	Use of apps in the general population	41%
Vera Cruz [25]	Use of apps for clinical purposes	18%
Bell et al. [26]	Use of apps for clinical purposes	Moderate reduction of anxiety and depression
Grand View Research [27]	Global market projection	Is projected to reach USD 17.5 billion by 2030

Table III below synthesizes the main digital public safety platforms, emergency response systems, and their territorial scope.

Main Platforms / Systems	Scope	Indicators
Life360	EE.UU.	More than 50 million active users.
Citizen	EE.UU.	Over 5 million users; coverage in more than 60 cities.
Next Generation 911 (NG911)	EE.UU.	Uneven implementation across states.
PSAP (EENA)	More than 60 countries	Transition toward next-generation models.
GAO	EE.UU.	Financial, technical, and institutional limitations.

B. Actual Scope in the Field of Public Safety

In the field of public safety, various indicators show sustained growth of community-based digital platforms. Among them, Life360 reports more than 50 million active users, with presence in approximately one out of nine families in the United States, according to company data (Life360, 2023).

Another relevant case is the Citizen application, designed for real-time incident notifications, which recorded more than 5 million active users in 2020 and was present in over 60 U.S. cities, according to Bradbury (2025). However, these data must be interpreted with caution, as they do not originate from official public safety agencies but rather from reports disseminated by the financial-technology sector. This suggests primarily market growth and increases social visibility of these platforms, rather than direct validation of their impact on safety.

On this matter, although the expansion of these technologies may be positive as a complementary tool to support public safety, there is still no conclusive evidence demonstrating their effective integration into official emergency services such as 911. Nor is there solid proof that their widespread use directly increases perceptions of protection or objective levels of public safety.

DISCUSSION OF FINDINGS

The results of the review highlight a central paradox: while mobile applications demonstrate high functional efficacy and sustained growth in terms of adoption and market expansion, their social impact remains conditioned by profound inequalities in access, institutional trust, and systemic integration (Teke et al., 2025; Wen & Tian, 2024). In particular, empirical evidence from the United States shows that racial minorities, rural communities, and households with lower socioeconomic status face significant barriers to accessing and using these technologies, which limits their effective social coverage. This tension constrains their reach as universal devices of emotional and physical protection.

Concerning this, the literature agrees that the consolidation of these technologies requires institutional validation strategies, through their formal incorporation into public policies, school programs, primary healthcare services, and municipal security plans (McCloud et al., 2020; Hernandez & Roberts, 2018). Data on mental health applications indicate

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.05

that, despite high smartphone ownership, only a limited percentage of individuals with diagnosed disorders access and sustain their use, underscoring the need for interventions that combine institutional recommendation, digital education, and continuous support. Official endorsement by governments, hospitals, and educational systems emerges as a key factor in strengthening social trust and promoting sustained use.

Another critical axis is interoperability with public services. Integration with emergency hotlines, healthcare centers, social assistance networks, and law enforcement agencies would allow platforms such as Life360 and Citizen to evolve from isolated tools into genuine digital infrastructures of social protection (EENA, 2023; United States Government, 2025). However, it should be noted that penetration data for Life360 and Citizen come primarily from commercial reports, which limits the possibility of drawing definitive conclusions about their actual operational impact. Strategic social integration, through mechanisms of community participation and digital support networks, could foster their appropriation without falling into addictive dynamics (Elphick et al., 2021).

Regarding future trends (2025–2030), the field of mental health is moving toward intelligent personalization, driven by emotional AI, convergence with wearable devices, and the consolidation of hybrid therapy models combining human professionals and automated systems (Luxton et al., 2016; Hwang et al., 2021). However, the effectiveness of these innovations will depend on overcoming structural barriers of access, digital literacy, and socioeconomic inequality, so that benefits are not concentrated exclusively in privileged segments. In parallel, the public safety sector is advancing toward predictive alert systems, interoperability with urban services, and the strengthening of neighborhood collective intelligence (Life360, 2023; Bradbury, 2025). Yet, evidence of direct operational impact remains limited and heterogeneous, requiring longitudinal and comparative evaluations to validate the real benefits in public safety.

At the same time, the advancement of these technologies intensifies debates on surveillance, algorithmic bias, and data governance (Kitchin, 2014; WHO, 2022). Ethical regulation and institutional responsibility thus emerge as indispensable conditions for these applications to consolidate as legitimate, inclusive, and socially sustainable tools (Mihale-Wilson et al., 2025).

Within this framework, the reviewed literature emphasizes that the social sustainability of these platforms depends not only on data protection and anonymization, but also on strategies of community participation and inclusive design that take into account cultural, linguistic, and socioeconomic differences.

CONCLUSIONS

This study, based on a narrative and documentary review of scientific literature, institutional reports, and representative digital platforms, aimed to analyze the adoption, impact, and challenges of mobile applications in the fields of mental health and public safety, with particular emphasis on the United States context. This methodology allowed for the identification of usage patterns, structural gaps, operational limitations, and opportunities for institutional integration of these technologies.

The findings demonstrate that mobile applications constitute a critical digital infrastructure for the comprehensive protection of citizens, articulating technology, community participation, and social inclusion potential. However, their effective social impact is conditioned by inequalities in access, digital literacy, institutional trust, and fragmented interoperability with public systems. Racial minorities, rural communities, and households with lower socioeconomic status face significant barriers to fully benefiting from these tools, limiting their reach as universal devices of emotional and physical protection.

The review shows that the consolidation of these technologies requires institutional validation, through their incorporation into public policies, educational programs, primary healthcare services, and municipal security plans. Likewise, interoperability with public services, community participation, and strategic social integration are identified as key factors to transform these platforms from isolated tools into sustainable and socially inclusive digital infrastructures.

Regarding future trends (2025–2030), the field of mental health is oriented toward intelligent personalization, convergence with wearable devices, and hybrid therapy models, while public safety is advancing toward predictive alerts, urban interoperability, and the strengthening of neighborhood collective intelligence. Nevertheless, the effectiveness of these innovations will depend on overcoming existing structural barriers and ensuring ethical regulation, data protection, and institutional responsibility, to prevent benefits from being concentrated exclusively among privileged segments.

In summary, mobile applications should not be considered merely technological products, but rather vehicles of emotional resilience, situational prevention, and collaborative governance. Their social and functional sustainability requires comprehensive strategies that combine inclusive design, institutional validation, technical interoperability, and digital inclusion policies, enabling their benefits to be distributed equitably and contributing to the digital transformation of social protection.

USE OF ARTIFICIAL INTELLIGENCE TOOLS

The author declares that during the preparation of this manuscript, an artificial intelligence-based tool was used exclusively for translation. This tool was employed to improve linguistic quality and readability, while the scientific content, data analysis, interpretations, and conclusions are the sole responsibility of the author.

REFERENCES

[1] C. Ventola, "Mobile devices and apps for health care professionals: uses and benefits," *P&T*, vol. 39, no. 5, pp. 356–364, May 2014. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/24883008/.

[2] D. D. Luxton, E.-L. Nelson, and M. M. Maheu, *A Practitioner's Guide to Telemental Health: How to Conduct Legal, Ethical, and Evidence-Based Telepractice*. Washington, DC: American Psychological Association, 2016. doi: 10.1037/14938-000.

[3] W. J. Hwang, J. S. Ha, and M. J. Kim, "Research trends on mobile mental health application for general population: a scoping review," *Int. J. Environ. Res. Public Health*, vol. 18, no. 5, p. 2459, Mar. 2021. doi: 10.3390/ijerph18052459.

[4] M. L. Este and B. C. Havard, "Mental health mobile apps: from infusion to diffusion in the mental health social system," *JMIR

Ment. Health*, vol. 2, no. 1, p. e10, Mar. 2015. doi: 10.2196/mental.3954.

[5] R. Kitchin, *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. London, U.K.: SAGE Publications Ltd, 2014.

[6] World Health Organization, *World Mental Health Report: Transforming Mental Health for All*. Geneva, Switzerland: WHO, 2022. [Online]. Available: https://iris.who.int/server/api/core/bitstreams/40e5a13a-fe50-4efa-b56d-6e8cf00d5bfa/content.

[7] T. McCloud, R. Jones, G. Lewis, V. Campana, and E. Tsakanikos, "Efficacy of a mobile application intervention for anxiety and depression symptoms in university students: randomized controlled trial," *JMIR Mhealth Uhealth*, vol. 8, no. 7, p. e15418, Jul. 2020. doi: 10.2196/15418.

[8] J. Firth, J. Torous, J. Nicholas, R. Carney, A. Pratap, S. Rosenbaum, and J. Sarris, "The efficacy of smartphone-based mental health interventions for depressive symptoms: a meta-analysis of randomized controlled trials," *World Psychiatry*, vol. 16, no. 3, pp. 287–298, Oct. 2017. doi: 10.1002/wps.20472.

[9] A. Almuqrin, R. Hammoud, I. Terbagou, S. Tognin, and A. Mechelli, "Smartphone apps for mental health: systematic review of the literature and five recommendations for clinical translation," *BMJ Open*, vol. 15, no. 2, p. e093932, Feb. 2025. doi: 10.1136/bmjopen-2024-093932.

[10] R. Kitchin, "The real-time city? Big data and smart urbanism," *GeoJournal*, vol. 79, pp. 1–14, 2014. doi: 10.1007/s10708-013-9516-8.

[11] D. Lupton, *Digital Sociology*. Cambridge, U.K.: Polity Press, 2015. [Online]. Available: https://www.dhi.ac.uk/san/waysofbeing/data/data-crone-lupton-2015b.pdf.

[12] S. Berrouiguet, E. Baca-García, S. Brandt, M. Walter, and P. Courtet, "Fundamentals for future mobile-health (mHealth): a systematic review of mobile phone and web-based text messaging in mental health," *J. Med. Internet Res.*, vol. 18, no. 6, p. e135, Jun. 2016. doi: 10.2196/jmir.5066.

[13] J. M. Lipschitz, R. Van Boxtel, J. Torous, J. Firth, J. G. Lebovitz, K. E. Burdick, and T. P. Hogan, "Digital mental health interventions for depression: scoping review of user engagement," *J. Med. Internet Res.*, vol. 24, no. 10, p. e39204, Oct. 2022. doi: 10.2196/39204.

[14] S. Lehtimäki, J. Martic, B. Wahl, K. T. Foster, and N. Schwalbe, "Evidence on digital mental health interventions for adolescents and young people: systematic overview," *JMIR Ment. Health*, vol. 8, no. 4, p. e25847, Apr. 2021. doi: 10.2196/25847.

[15] D. Lyon, *The Culture of Surveillance: Watching as a Way of Life*. Cambridge, U.K.: Polity Press, 2018.

[16] B. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, and L. Floridi, "The ethics of algorithms: mapping the debate," *Big Data & Society*, vol. 3, no. 2, Nov. 2016. doi: 10.1177/2053951716679679.

[17] J. Teke, D. B. Olawade, N. Leena, K. Weerasinghe, S. Mc Lernon, and C. Moorley, "Digital health disparities: a review of barriers and solutions for racially diverse groups," *Int. J. Med. Inform.*, vol. 206, p. 106173, Feb. 2026. doi: 10.1016/j.ijmedinf.2026.106173.

[18] K. Hernandez and T. Roberts, "Leaving no one behind in a digital world," K4D Emerging Issues Report, Institute of Development Studies, Brighton, U.K., 2018. [Online]. Available: https://assets.publishing.service.gov.uk/media/5c178371ed915d0b8a31a404/Emerging_Issues_LNOBDW_final.pdf.

[19] X. Wen and B. Tian, "How does digital integration influence the mental health of low-income populations?" *Healthcare*, vol. 12, no. 24, p. 2593, Dec. 2024. doi: 10.3390/healthcare12242593.

[20] C. Elphick, R. Philpot, M. Zhang, A. Stuart, Z. Walkington, and L. A. Frumkin, "Building trust in digital policing: a scoping review of community policing apps," *Police Practice and Research*, vol. 22, no. 5, pp. 1469–1491, 2021. doi: 10.1080/15614263.2020.1861449.

[21] C. Mihale-Wilson, P. Felka, and O. Hinz, "Mobile ICT outages and public safety: is there a digital divide in terms of safety?" *Eur. J. Inf. Syst.*, pp. 1–20, 2025. doi: 10.1080/0960085X.2025.2523984.

[22] M. Reinert, D. Fritze, and T. Nguyen, "The state of mental health in America 2023," Mental Health America, Alexandria, VA, Oct. 2022. [Online]. Available: https://mhanational.org/wp-content/uploads/2024/12/2023-State-of-Mental-Health-in-America-Report.pdf.

[23] Y. D. Guracho, S. J. Thomas, and K. T. Win, "Smartphone application usage patterns for mental disorders: systematic review and meta-analysis," *Int. J. Med. Inform.*, vol. 179, p. 105217, Nov. 2023. doi: 10.1016/j.ijmedinf.2023.105217.

[24] S. Fürtjes, E. Gebel, H. Kische, and K. Beesdo-Baum, "Characteristics of mental health app usage: a cross-sectional survey in the general population," *BMC Public Health*, vol. 24, no. 1, p. 3133, Nov. 2024. doi: 10.1186/s12889-024-20500-1.

[25] G. Vera Cruz, E. Aboujaoude, R. Khan, L. Rochat, F. Ben Brahim, R. Courtois, and Y. Khazaal, "Smartphone apps for mental health and wellbeing: a usage survey and machine learning analysis of psychological and behavioral predictors," *Digit. Health*, vol. 9, p. 20552076231152164, Jan. 2023. doi: 10.1177/20552076231152164.

[26] I. H. Bell, A. Thompson, L. Valentine, S. Adams, M. Alvarez-Jimenez, and J. Nicholas, "Ownership, use of, and interest in digital mental health technologies among clinicians and young people across a spectrum of clinical care needs: cross-sectional survey," *JMIR Ment. Health*, vol. 9, no. 5, p. e30716, May 2022. doi: 10.2196/30716.

[27] Grand View Research, "Mental health apps market size to reach $17.52 billion by 2030," Report, San Francisco, CA, USA, 2025. [Online]. Available: https://www.grandviewresearch.com/press-release/global-mental-health-apps-market.

[28] K. Collura, "Life360 surpasses 50 million monthly active users," *PR Newswire*, Mar. 2023. [Online]. Available: https://www.prnewswire.com/news-releases/life360-surpasses-50-million-monthly-active-users-301779366.html.

[29] R. Bradbury, "Crime safety app Citizen targets $42M in fundraising push," *PitchBook News*, Jan. 2025. [Online]. Available: https://pitchbook.com/news/articles/crime-safety-app-citizen-targets-42-million.

[30] European Emergency Number Association (EENA) and Government Accountability Office (GAO), "Public safety answering points: global edition 2024," Feb. 2025. [Online]. Available: https://eena.org/wp-content/uploads/2024_PSAPs_Global_Edition_v02_Abstract_v02.pdf.

[31] United States Government, "Next Generation 911," *911.gov*, 2025. [Online]. Available: https://www.911.gov/issues/ng911/.

[32] U.S. Government Accountability Office (GAO), "Next Generation 911: Some federal agencies have begun planning, but few have upgraded their call centers," Report to Congressional Committees, GAO-24-106783, 2024. [Online]. Available: https://www.gao.gov/assets/gao-24-106783.pdf.

AUTHORS

Diego Mattera

Diego Mattera, in terms of academic training, holds a Bachelor's degree in Criminalistics and a Master's degree in Criminology, Delinquency, and Victimology from the Universidad Internacional de Valencia, together with additional postgraduate studies in Cybersecurity, Criminal Profiling, and Cybercrime Investigation. He has extensive experience in law enforcement, complex crime investigation, and fraud prevention. He began his career within the Argentine security forces, where he managed criminal complaints, ethical violations, and forensic procedures such as fingerprint extraction. Subsequently, at the
Interpol headquarters, he carried out complex investigations into organized crime, employing intelligence tools, databases, and criminal profiling methodologies. His work included the preparation of comprehensive reports and coordination with international policing bodies. He currently works as a researcher in the field of Fraud Prevention in the private sector, where he analyzes cases of fraud, theft, and losses through advanced systems such as TMS/WMS and satellite monitoring, while also advising on preventive strategies for logistics operations.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026

IoT-Enabled Deep Reinforcement Learning for Adaptive Waste Management in Hospital Environments

ARTICLE HISTORY
Received 24 December 2025
Accepted 23 June 2026
Published 7 July 2026

Muhammad Masood Ul Rahman Usmani
Department of Computer Science
COMSATS University Islamabad, Sahiwal Campus
Sahiwal, Pakistan
muhammadmasoodulrahmanusmani@gmail.com
ORCID: 0009-0001-9948-111X

Rimsha Rafiq
Department of Computer Science
COMSATS University Islamabad, Sahiwal Campus
Sahiwal, Pakistan
rimshach026@gmail.com
ORCID: 0009-0009-4002-0798

Makki Riaz Khan
Department of Information Technology
Bahauddin Zakariya University, Lodhran Campus
Lodhran, Pakistan
makkiriazkhanwapda@gmail.com
ORCID: 0009-0009-0698-3857

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026

M. M. U. R. Usmani, R. Rafiq, and M. R. Khan, "IoT-Enabled Deep Reinforcement Learning for Adaptive Waste Management in Hospital Environments", Latin-American Journal of Computing (LAJC), vol. 13, no. 2, 2026.

IoT-Enabled Deep Reinforcement Learning for Adaptive Waste Management in Hospital Environments

Muhammad Masood Ul Rahman Usmani
COMSATS University Islamabad, Sahiwal Campus
Department of Computer Science
Sahiwal, Pakistan
muhammadmasoodulrahmanusmani@gmail.com

Rimsha Rafiq
COMSATS University Islamabad, Sahiwal Campus
Department of Computer Science
Sahiwal, Pakistan
rimshach026@gmail.com

Makki Riaz Khan
Bahauddin Zakariya University, Lodhran Campus
Department of Information Technology
Lodhran, Pakistan
makkiriazkhanwapda@gmail.com

***Abstract*—Healthcare industry produces considerable amount of biomedical waste, which needs efficient and safe ways of handling. The research paper puts forward a scheme of IoT-enabled DRL for hospital waste management. Unlike previously known schemes, the suggested scheme incorporates IoT-based sensing capabilities with a Proximal Policy Optimization agent. The performance of the framework is analyzed by developing a custom simulation environment based on the OpenAI Gym library, wherein the process of waste production is modeled as stochastic. According to the experimental results, the suggested scheme of PPO surpasses its competitors in all key criteria. Statistical validation using ANOVA and t-tests confirms that the improvements are significant. The findings highlight the potential of IoT-DRL integration for intelligent, adaptive, and efficient hospital waste management systems.**

***Keywords*—Biomedical Waste Management, Internet of Things (IoT), Deep Reinforcement Learning (DRL), Healthcare Systems, Intelligent Optimization, Hospital Safety, Sustainability.**

I. Introduction

The rising number of healthcare institutions has worsened the situation regarding biomedical and hospital waste management. Inadequate methods of separating and disposing of hospital waste can pose severe threats to the environment and raise the risks of disease transmission to healthcare providers and communities nearby [1], [2]. Conventional separation by healthcare providers can be laborious, prone to mistakes, and dangerous to healthcare workers [3], [4].

Recent developments in IoT and AI technologies have opened avenues towards new paradigms in smart waste management. Low-cost technologies in IoT (e.g., ultrasonic level sensors, RFID tags, and environment sensors) enable waste level detection at all times and provide real-time input to central processing units [5]–[7]. The application of intelligent systems to already existing infrastructure in IoT makes it possible to manage waste proactively and decrease the chances of overflow and pollution in sensitive sectors like hospitals [7].

Machine learning algorithms, specifically deep learning networks such as EfficientNet, ResNeXt, MobileNet, and YOLO, have achieved significant success in classifying different types of healthcare waste accurately with a classification accuracy of more than 95% [1], [4], [8]. Such systems bring the process of proper healthcare waste segregation at par with the national healthcare waste management guidelines, thus preventing the risks of exposure to harmful healthcare waste [8], [22].

However, despite such advancements, the majority of existing systems are still either reactive or limited to a labscale environment. For instance, even though IoT-enabled systems can send notifications once the bin is full, they do not have the predictive models required for optimal collection route planning [7], [19]. Similarly, vision-based sorting models, although data-intensive, may fail to effectively adapt to a healthcare setting [9], [10]. Thus, a critical need emerges for adaptive and smart systems to adapt to the ever-changing hospital waste production dynamics.

Deep Reinforcement Learning (DRL) is also an emerging promising technology for adaptive decision-making in challenging and uncertain environments. In contrast to static ML models, DRL agents are capable of learning optimal waste management techniques such as dynamic routing, vehicle allocation, and prioritization of bins through trial and error interactions in the environment [11], [12]. Multi-agent DRL techniques are also further extended and employed for simultaneous allocation of waste collection vehicles that aim to reduce energy costs, minimize delays in waste collection, and avoid the formation of dangerous wastes [11], [13].

In a hospital setting, the adoption of IoT-capable DRL models will be beneficial in making hospital waste management activities more predictive and adaptive. Real-time data inputs from sensors, along with models for reinforcement learning, such as Proximal Policy Optimization (PPO) and Deep Q-Networks (DQN), will be useful in making hospital waste management more adaptive by optimizing routes, minimizing overflow levels, and maintaining biomedical waste regulations [7], [11].

This study introduces an IoT-assisted Deep Reinforcement

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.06

Learning framework for adaptive hospital waste management. The framework is founded on the recent advancements in the classification of healthcare waste in [8], proactive IoT assistance in hospital waste management in [7], and multiagent reinforcement learning with the goal of smart cities in [11].

The contributions of this paper are:

- An IoT-based sensor system capable of continuously tracking waste generation, availability of bins, and other risk factors in hospital environments.
- An accurate model for classification of different types of medical waste (hazardous, infectious, general, sharps, etc.) using deep learning techniques on images and sensor data.
- A DRL agent responsible for learning adaptive policies regarding waste collection, bin services, routing, and disinfection or storage considering safety and cost aspects.
- Evaluation on hospital data to assess improved classification accuracy, lower cost and risk, and ability to adapt to variable amounts of waste.

The second part highlights related literature on IoT, machine learning, and RL on Waste Management. The third section reviews the proposed IoT-enabled DRL framework. The fourth part highlights the experiment setup. The fifth section presents results discussion and comparison. The sixth part provides conclusions and future directions.

II. Related Work

Hospital and biomedical waste management has been an increasingly growing problem has attracted a lot of interest from the scientific community lately. This is because of the associated effects of infection control, sustainability, and efficiency. Traditional methods involving manual segregation and collection at pre-documented times have been found to be increasingly ineffective in dealing with the dynamic and dangerous nature of hospital waste [1] [2]. As such, there has been considerable interest in the use of emerging technologies such as IoT, ML, and DRL for smart and automated hospital waste management models.

A. Deep Learning Applied to Healthcare Waste Classification

A recent area that has shown great promise for automatic classification of healthcare waste is based on the principles of deep learning. Various research articles that applied transfer learning with ResNeXt, EfficientNet, and MobileNet models have demonstrated classification achievements above 90% for different categories of healthcare wastes [1] [2]. Specifically, YOLO models have also demonstrated their efficiency with respect to real-time processing, with YOLOv5 and YOLOv8 models achieving above 95% classification accuracy for different categories of wastes with lower inference times, making these models suitable for hospital settings [3] [4] [8]. Nevertheless, these models also rely heavily upon large-scale healthcare datasets, and their applicability for classification with hospital settings has shown some challenges [9] [10].

B. IoT-Enabled Smart Waste Systems

IoT-Smart garbage management systems in the IoT environment are also explored in terms of urban waste collection points (WCC) and smart cities. The usage of low-cost sensor nodes and cloud computing contributes to realtime monitoring of the fill level, temperature, humidity, and environmental aspects of garbage bins in the IoT-based system [5]– [7]. ProWaste technology, which involves the application of machine learning algorithms and IoT sensors, demonstrates prediction rates of above 99% in forecasting situations of overflow at WCCs. The conceptual structures of smart garbage bins in the IoT environment are also explored, incorporating the use of sensors, compression systems, and solar-based circuits to maximize effectiveness and minimize costs [7] [11] [20] [21].

C. Reinforcement Learning and Optimization

Routing optimization for waste collection is an important applications area in the healthcare sector, considering the consequences of inefficient waste collection, which can result in accelerated health risk escalation. Conventional methods for optimization, also originating in the Vehicle Routing Problem, have been adapted for superior effectiveness, thereby incorporating stochastic optimization techniques [12] [19]. In the more recent past, reinforcement learning has been employed for effective dynamic optimization in uncertain settings for waste routing, thereby facilitating adaptability according to real-time waste generation levels [11] [13]. Furthermore, multi-agent deep reinforcement learning setups ensure greater flexibility, incorporating vehicle dispatch and management for multiple zones in an upscale healthcare facility, thereby increasing energy efficiency and overall [11].

D. Integration in Healthcare Contexts

However, their application to a hospital environment is still very limited despite significant advances in general smart waste management. Biomedical waste is unique in hazardous material classification, strict regulatory compliance, and infection risks for healthcare staff [8] [22]. IoT-enabled sensing with DRL can help build adaptive, predictive frameworks that manage the complexity of the waste system of a hospital. In addition, such integration fits into the sustainability and public health goals while ensuring compliance with the national and international biomedical waste standards [2], [7], [11].

Summarily, from the literature, there are promising fronts but also some identified gaps. The existing systems based on IoT are mostly on notification, and classifications based on deep learning focused on correctness in lieu of optimized routes. The application of reinforcement learning, although a robust method, is less considered in a healthcare setting scenario. There consequently arises a need for a unified framework of DRL using IoT-enabled adaptive hospital waste management.

Table I summarizes recent approaches in IoT, machine learning, and DRL-based waste management systems. The comparison shows the limitations of the existing methods, motivating the need for an integrated IoT-DRL framework.

III. Methodology

The proposed framework utilizes a combination of IoT sensing and DRL optimization techniques to provide an adaptive biomedical waste management system within a hospital setting. The methodology is comprised of three fundamental

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.06

TABLE I. Comparative Summary of Recent IoT, ML/DL, and DRL Approaches for Waste Management (2021–2025)

Authors / Year	Method / Model	Application / Dataset	Key Results / Findings
Zhou et al. (2022) [1]	ResNeXt-50 (DL)	Private 8-class medical waste dataset	Achieved 97.2% accuracy in healthcare waste classification.
Kumar et al. (2021) [2]	EfficientNet-B7 (TL, DL)	COVID-related biomedical waste streams	Reported 99% accuracy; highlighted AI for circular economy in healthcare waste.
Kunwar & Rai (2025) [8]	YOLOv5-s, YOLOv8, EfficientNet-B0	Medical Waste Dataset 4.0 + Pharma-biomedical dataset (Nepal)	YOLOv5-s achieved 95.06% accuracy; deployed with bin-color mapping to Nepal's HCW standards.
Mok (2024) [4]	YOLO + IoT Integration	67,860 images of HCW detection	Achieved mAP of 98%; demonstrated IoT-enhanced automated sorting.
Lahoti et al. (2024) [3]	Computer Vision + Robotic Arm	Multi-class waste segregation prototype	Enabled real-time robotic segregation of hospital waste.
Stephan et al. (2025) [6]	ProWaste (IoT + ML, Decision Tree + BPSO)	Urban WCCs, 6,954 daily records (Bengaluru)	Reduced missed pickups; >99.8% macro-F1 with only 3 predictive features; deployed via mobile app.
Patil et al. (2021) [5]	Ultrasonic Sensors + IoT	E-waste and bin monitoring prototype	IoT-based notification system reduced manual inspections.
Shanthini et al. (2021) [7]	IoT, RFID, WSN	Smart City Waste Collection	Found LoRaWAN-based IoT systems outperform others in efficiency.
Rajani et al. (2022) [11]	Multi-Agent Deep RL (DRL)	IoT-driven smart waste management (simulation)	Proposed platform-agnostic DRL framework; optimized routing, reduced overflow.
Mishra et al. (2022) [12]	Route Optimization (VRP + IoT)	Bhubaneswar City MSW data	Reduced vehicle distance by 30.28%, OpEx by 29.07%.
Abuga et al. (2022) [13]	IoT + Fuzzy Logic	Real-time smart garbage bins	Achieved high reliability in dynamic bin monitoring and waste-level prediction.
Gondal et al. (2023) [9]	Hybrid Deep Learning Model	Real-time garbage classification	Achieved 99% training/validation accuracy with automated bin sorting.
Zhang et al. (2022) [10]	Cascade R-CNN (enhanced with dilated convolutions)	Garbage detection for small objects	Improved precision in detecting small waste objects in cluttered environments.

components, namely: (i) IoT Layer, (ii) DRL Layer, and (iii) System Workflow. Figure 1 shows the overall architecture.

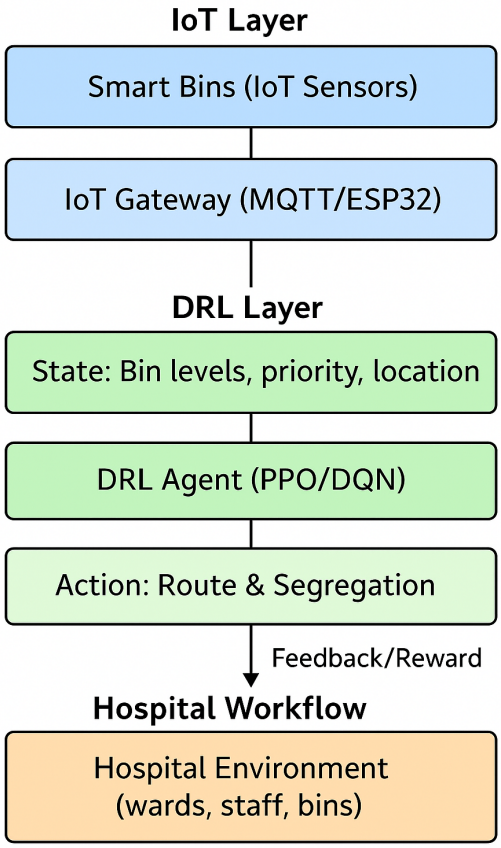


Fig. 1: Proposed IoT-Enabled DRL Framework for Adaptive Hospital Waste Management

A. *IoT Layer*

This layer contains a network of intelligent biomedical waste dumpsters with ultrasonic sensors for measuring the fill levels of the dumpsters, an RFID for identifying biowaste types, and temperature sensors for the detection of infectious diseases. Data from the dumpers is relayed through lightweight messaging transports such as MQTT to a cloud server. This enables real-time mapping of biowaste generation behavior in hospital wards, operation theatres, and labs. Figure 2 illustrates the PPO architecture used in this study.

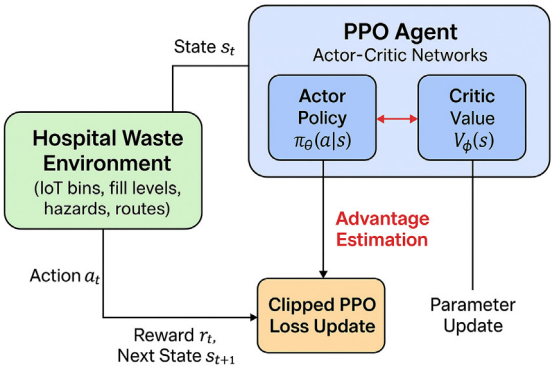


Fig. 2: Proposed PPO Model Architecture

B. *DRL Layer*

The DRL layer formulates waste management as a sequential decision-making problem. Every individual in charge of collecting waste material from different hospital locations (for instance, a robot cart or human personnel) is considered an agent functioning within the hospital premises. The state space will consist of bin levels, location, waste importance

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.06

Algorithm 1 Training Procedure of the Proposed PPO Agent

Randomly initialize the policy parameters θ and value network parameters ϕ
for each training episode **do**
  Reset the environment and obtain the initial state s
  for each interaction step until horizon T **do**
    Sample an action a according to the current policy $\pi_\theta(\cdot|s)$
    Apply a to the environment
    Receive reward r and the subsequent state s'
    Save the transition (s, a, r, s') in the rollout buffer
    Set $s \leftarrow s'$
  end for
  Estimate discounted returns and compute advantages using GAE
  Evaluate the clipped objective function:

$$L_{\text{PPO}} = \mathbb{E}\Big[\min\Big(r(\theta)\hat{A}, \text{clip}(r(\theta), 1-\epsilon, 1+\epsilon)\hat{A}\Big)\Big]$$

  Optimize the policy network using gradient ascent
  Update the value network by minimizing the value prediction error
end for

(hazardous or non-hazardous), and availability of employees. Actions will comprise routing actions (next bin to collect, segregation of waste material, or inaction). Reward functions will aim at (i) reducing the routing distance and duration of waste collection; (ii) avoiding overflowing of bins containing hazardous waste materials; and (iii) compliance with the biomedical safety standards. PPO will be chosen as the algorithm to use based on DRL since it is stable and suited for continuous state spaces.

C. System Workflow

The overall system workflow is as follows:

- IoT sensors collect real-time data from hospital waste bins.
- Data is transmitted to a cloud server and preprocessed (normalization, bin classification).
- The DRL agent receives the current hospital state and selects an optimal action (e.g., which bin to service).
- The action is executed by the waste collection system (robot/human).
- The environment returns feedback (updated state, reward signal).
- The DRL model updates its policy iteratively through training episodes.

This adaptive loop allows the system to dynamically optimize collection routes and schedules while minimizing risks of contamination and operational inefficiencies.

1) PPO Training Algorithm: Use Proximal Policy Optimization (PPO) to train the DRL agent [14]. Algorithm 1 presents the pseudocode of the training process.

D. Simulation Environment

For assessing the proposed framework, a simulation environment was designed with the help of the OpenAI Gym API. The environment considers hospital waste bins, collectors,

Algorithm 2 Custom OpenAI Gym Environment for Hospital Waste Management

Require: Number of bins B, Hospital layout L, Maximum steps T
Ensure: State transitions, reward signals
Initialize: Bin fill levels $f_b = 0$, Waste types $w_b \in \{hazardous, non\text{-}hazardous\}$, Agent position p_0
for Epoches $= 1$ to N **do**
  Reset: $f_b \sim U(0, 0.3)$, $p_0 =$ nurse station
  for $t = 0$ to $T - 1$ **do**
    State: $s_t = \{f_b, w_b, p_t\}$
    Agent selects action $a_t \in$ {move to bin, collect, idle}
    if action == collect **then**
      Empty selected bin, update $f_b \leftarrow 0$
      Reward $r_t = +\alpha$ (hazardous) or $+\beta$ (non-hazardous)
    end if
    Update fill levels: $f_b \leftarrow f_b + \delta$
    if $f_b > 1.0$ **then**
      Penalize: $r_t = -\gamma$ (overflow)
    end if
    Update agent position p_{t+1} and hospital state
  end for
  if all bins empty or $t = T$ **then**
    Terminate episode
  end if
  Return trajectory $\{s_t, a_t, r_t, s_{t+1}\}$ for PPO training
end for

and IoT sensors dynamics. In each episode, the simulation reflects the dynamics of a hospital shift, where the agent tries to find an optimal path and waste segregation strategy. The algorithm for the environment is presented in Algorithm 2.

IV. Results & Experiments

This section outlines the experimental configuration, explains the evaluation criteria, and analyzes the performance of IoT-based DRL framework for adaptive hospital waste management system. The experiments have been performed using the Python environment that uses the custom OpenAI Gym environment described in Algorithm 2 and TensorFlow for DRL. Experiments were conducted on a workstation with Intel i7-12700 CPU, 32GB RAM, and NVIDIA RTX 3080 GPU.

A. Experimental Setup

The hospital environment was simulated with $B = 50$ bins distributed across wards, laboratories, and operation theaters. Each bin was randomly assigned as hazardous (30%) or non-hazardous (70%), and waste generation rates followed a Poisson distribution. It is important to note that the experimental evaluation is conducted using a simulated hospital environment. IoT data streams are synthetically generated to emulate real-world waste generation patterns. The DRL agent (PPO) was compared against baseline methods:

- Rule-Based Scheduling (RBS): Fixed-time collection intervals.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.06

TABLE II. Simulation and PPO Hyperparameters

Parameter	Value
α (Hazardous reward)	10
β (Non-hazardous reward)	5
γ (Overflow penalty)	15
λ (Poisson rate)	0.2–0.5
Learning Rate	3×10^{-4}
Discount Factor (γ)	0.99
PPO Clipping (ϵ)	0.2
Batch Size	64
Epochs	10

TABLE III. Performance Comparison of Waste Management Approaches

Method	Overflow (%)	Coll. Time (min)	Haz. Score	Energy (kWh)
RBS	18.2	12.5	0.62	4.8
SPH	12.9	9.4	0.71	4.1
DQN	9.7	8.1	0.76	3.7
Proposed PPO	**5.4**	**6.9**	**0.89**	**3.2**

- Shortest Path Heuristic (SPH): Greedy routing to the nearest non-empty bin.
- Deep Q-Network (DQN): Model-free baseline with discrete actions.

B. Hyperparameter Settings

To ensure reproducibility, the following hyperparameters were used:

C. Metrics for Evaluation

The system was evaluated using the following metrics:

- **Overflow Rate (%)**: Percentage of bins that exceeded capacity.
- **Average Collection Time (min)**: Mean time per bin service.
- **Hazardous Waste Priority Score**: Ratio of hazardous waste collected before overflow.
- **Energy Efficiency (kWh)**: Energy consumed by collection robots/vehicles.

D. Results and Discussion

Table III shows the comparison across baseline methods. The proposed PPO-based DRL outperformed all baselines, achieving the lowest overflow rate and highest hazardous waste priority compliance. Fig. 3 and Fig. 4 illustrate the overflow reduction and training convergence, respectively.

The findings validate that DRL-based adaptive scheduling helps minimize the risks of overflow as well as enhances the efficiency of hazardous waste management. Additionally, the reduced energy consumption indicates the sustainability advantages of such an approach in practical settings of hospitals as well.

E. Statistical Significance Analysis

We further conducted statistical significance tests to assess the robustness of the outcomes of the proposed method and the baselines. We executed 30 independent episodes for each method based on randomly generated patterns of waste production. A one-way ANOVA test is followed by pairwise two-tailed t-tests using Bonferroni correction for overflow rate.

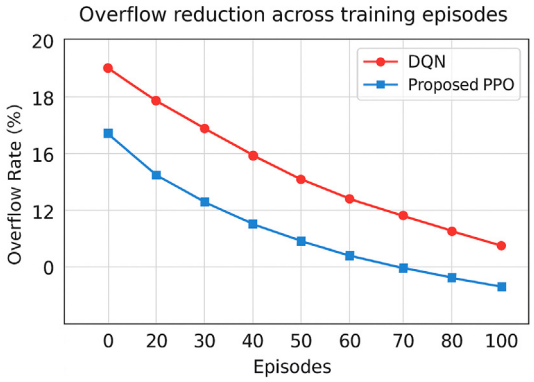


Fig. 3: Overflow reduction across training episodes

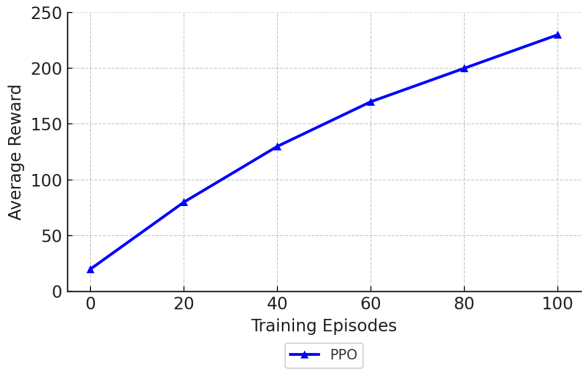


Fig. 4: Training convergence of PPO agent

Table IV shows that the ANOVA test yielded a statistically significant difference among methods ($p < 0.001$). Post-hoc pairwise t-tests (Table V) confirmed that the proposed PPO significantly outperformed RBS, SPH, and DQN.

TABLE IV. One-Way ANOVA Results for Overflow Rate Across Methods

Source	DF	F-Value	p-Value
Between Groups	3	24.56	< 0.001
Within Groups	116	–	–
Total	119	–	–

TABLE V. Pairwise t-Test Results (Overflow Rate, Bonferroni Corrected)

Comparison	t-Statistic	p-Value
PPO vs. RBS	7.82	< 0.001
PPO vs. SPH	6.11	< 0.001
PPO vs. DQN	3.45	0.002

The statistical analysis confirms that the improvements of PPO are not due to random chance, but are significant at the 95% confidence level. Thus, the proposed DRL-based approach provides a robust and reliable optimization strategy for hospital waste management.

F. Complexity Analysis

To assess the computational feasibility of the proposed framework, we analyzed the time and space complexity of both the IoT data acquisition and the DRL training phases.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.06

1) IoT Communication Overhead: The IoT layer relies on MQTT-based communication for transmitting sensor data from B bins. Each message has an average payload size of $O(1)$ (bin fill level, type, timestamp). Thus, the per-step communication complexity is

$$O(B) \tag{1}$$

For a hospital with $B = 100$ bins, the per-second communication load remains within 50–100 KB, which is well within the capabilities of low-cost WiFi/LoRa gateways.

2) PPO Training Complexity: The PPO algorithm involves two key components: trajectory collection and policy updates. Let N be the number of steps per episode, E the number of episodes, and M the number of policy update iterations.

- **Trajectory Collection**: Each step requires state evaluation and action selection, with complexity $O(d)$, where d is the dimensionality of the state vector. Thus, trajectory collection over N steps per episode costs

$$O(Nd) \tag{2}$$

- **Policy Update**: PPO uses stochastic gradient descent on batches of size b over M iterations. Each forward-backward pass has complexity $O(\theta)$, where θ is the number of neural network parameters. Thus, update complexity is

$$O(Mb\theta) \tag{3}$$

Overall training complexity is therefore

$$O(END + EMb\theta) \tag{4}$$

3) Space Complexity: The memory footprint consists of: replay buffer $O(Nd)$, neural network parameters $O(\theta)$, and IoT data queue $O(B)$. Total space complexity is

$$O(Nd + \theta + B) \tag{5}$$

In practice, with $N = 1000$ steps, $E = 500$ episodes, $M = 10$ iterations, $b = 64$, and $\theta \approx 10^5$, training can be completed using a single NVIDIA RTX 3060 GPU in under 4 hours. The IoT communication overhead is negligible relative to network capacity. Thus, the framework is computationally efficient and feasible for real-world hospital environments.

V. Discussion

The obtained results confirm that the proposed IoT-enabled PPO framework provides a significant improvement in hospital waste management efficiency compared to baseline approaches, including Rule-Based Scheduling (RBS), Shortest Path Heuristic (SPH), and Deep Q-Network (DQN). Specifically, the proposed model achieves lower overflow rates, reduced collection times, and improved prioritization of hazardous waste. These improvements are statistically validated using ANOVA and post-hoc t-tests, confirming that the observed performance gains are not due to random variation.

A. Interpretation of Results

These performance gains can be explained by two major reasons. Firstly, the IoT-based data acquisition layer allows for the adaptive tracking of waste production dynamics and thus helps make the decisions in a timely fashion depending on current conditions in the hospital, which is more efficient than static or heuristic methods.

Secondly, the PPO method offers stable learning in stochastic environments thanks to the clipped surrogate objective function that does not allow for very large updates of the policy function during learning. This feature is especially important in hospitals due to the uncertainty in waste production and other constraints.

B. Limitations

However, despite all the positive results achieved, some limitations should be mentioned. First of all, the evaluation of the proposed solution takes place within a simulation setting wherein the patterns of waste generation are artificially generated. Even though these patterns are created with an intention to mirror realistic conditions at the hospital, they might not be sufficiently diverse enough.

Waste generation in real life may be subject to some unexpected factors like outbreaks of infectious diseases, changes in seasons, and the increase in patient flow, among others. In particular, during a large-scale outbreak, the volume and content of biomedical waste may change substantially. Hence, the evaluation of the system that takes place within a simulation is an idealistic case study.

C. Ethical and Privacy Considerations

This approach uses information from hospital waste, some of which might be confidential metadata, like timestamps that may inadvertently correspond to patient behavior. Compliance with data protection laws, HIPAA and GDPR, is a prerequisite before implementation of such an approach.

In addition, although automation decreases the exposure to harmful waste for people, it is vital to retain proper human control. System malfunctions, wrong classifications or any other circumstances can be dangerous. Thus, the ethical implementation of the system requires proper mechanisms of monitoring and fail-safes.

D. Generalizability and Future Deployment

Though the model has been designed with hospital waste management in mind, the suggested IoT-DRL model can be easily customized for any other type of waste management operations, such as intelligent cities and industries. The implementation of multi-agent reinforcement learning will help scale the model even more, as it will allow coordination among several collectors.

Further research should be directed at applying the model in practice through collaboration with the healthcare sector. Besides, introducing uncertainty analysis, anomaly detection, and edge computing is a good idea as well.

VI. Conclusion and Future Work

This paper provided a novel IoT based deep reinforcement learning (DRL) approach to adaptive hospital waste management. The proposed method consists of using an IoT based

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.06

sensor along with proximal policy optimization (PPO) for dynamic optimization of waste collection, route planning, and prioritization of hazardous wastes.

From the results of our experiments performed in the custom simulation environment built on top of the OpenAI Gym, we can see that the behavior of the proposed DRL method is significantly better than those of the baselines methods RBS, SPH, and DQN. Our model provides the results with lower overflow rate, lesser collection time, and better prioritization of hazardous wastes.

The primary contributions of this study are summarized as follows:

- Development of an integrated IoT-DRL framework for adaptive hospital waste management.
- Design of a custom simulation environment to model realistic waste generation and collection dynamics.
- Comprehensive performance evaluation demonstrating the effectiveness of PPO in complex and stochastic environments.
- Analysis of computational complexity and consideration of ethical and deployment-related aspects.

However, there are some limitations of this study. The assessment was conducted based on artificial datasets in a simulation setting, which does not completely account for real-life randomness in hospital operations. Hence, the obtained results correspond to an ideal case scenario.

The future works will be directed towards practical implementation of the proposed approach together with hospitals. Furthermore, the use of the multi-agent reinforcement learning may make the system capable of collecting waste cooperatively via several autonomous agents in hospital areas. Some additional improvements might include robustness enhancement via modeling of uncertainties, anomalies detection and failure-proofed IoT communications.

Additionally, the employment of energy-efficient DRL models and edge computing methods will enhance the scalability and application of the framework in limited-resource settings. Moreover, apart from being applied to hospitals, the proposed framework can potentially be used for other applications such as smart city waste management systems, industrial waste handling systems and other public health infrastructure.

Thus, the combination of IoT and DRL may lead to the development of an advanced waste management framework.

Funding statement

There was no outside support for this study.

Acknowledgment

The Department of Computer Science at COMSATS University Islamabad's Sahiwal Campus is acknowledged by the authors for providing the academic resources and assistance required to complete this study.

Author contributions

The authors' contributions follow the CRediT (Contributor Roles Taxonomy) as follows:

- **Muhammad Masood Ul Rahman Usmani**: Conceptualization, Methodology, Writing Original Draft & Editing
- **Rimsha Rafiq**: Data Curation, Visualization, Review & Editing
- **Makki Riaz Khan**: Conceptualization, Writing – Review & Editing

References

[1] H. Zhou, X. Yu, A. Alhaskawi, Y. Dong, Z. Wang, Q. Jin, X. Hu, Z. Liu, V. G. Kota, M. H. Abdulla, S. H. A. Ezzi, B. Qi, J. Li, B. Wang, J. Fang, and H. Lu, "A deep learning approach for medical waste classification," *Scientific Reports*, vol. 12, no. 1, p. 2159, 2022.

[2] N. M. Kumar, M. A. Mohammed, K. H. Abdulkareem, R. Damasevicius, S. A. Mostafa, M. S. Maashi, and S. S. Chopra, "Artificial intelligence-based solution for sorting COVID related medical waste streams and supporting data-driven decisions for smart circular economy practice," *Process Safety and Environmental Protection*, vol. 152, pp. 482–494, 2021.

[3] J. Lahoti, J. Sn, M. V. Krishna, M. Prasad, R. Bs, N. Mysore, and J. S. Nayak, "Multi-class waste segregation using computer vision and robotic arm," *PeerJ Computer Science*, vol. 10, p. e1957, 2024.

[4] M. H. Mok, "YOLO combined with IoT for detection of healthcare waste," *Applied Sciences*, vol. 14, no. 3, p. 1167, 2024.

[5] U. Patil *et al.*, "IoT based smart waste management system," in *Proc. IEEE Int. Conf. Inventive Research in Computing Applications (ICIRCA)*, 2021, pp. 1–6.

[6] T. Stephan, S. M. Hari Krishna, C.-C. Lin, U. Sumesh, S. Agarwal, and H. Kim, "ProWaste for proactive urban waste management using IoT and machine learning," *Scientific Reports*, vol. 15, no. 1, p. 27790, 2025.

[7] S. Vishnu, S. R. J. Ramson, S. Senith, T. Anagnostopoulos, A. M. AbuMahfouz, X. Fan, S. Srinivasan, and A. A. Kirubaraj, "IoT-enabled solid waste management in smart cities," *Smart Cities*, vol. 4, no. 3, pp. 1004–1017, 2021.

[8] S. Kunwar and P. Rai, "Healthcare waste classification using deep learning aligned with Nepal's bin color guidelines," *arXiv preprint arXiv:2508.07450*, 2025.

[9] A. U. Gondal, M. I. Sadiq, T. Ali, M. Irfan, A. Shaf, M. Aamir, M. Shoaib, A. Glowacz, R. Tadeusiewicz, and E. Kantoch, "Real time multipurpose smart waste classification model for efficient recycling in smart cities using multilayer convolutional neural network and perceptron," *Sensors*, vol. 21, no. 14, p. 4916, 2021.

[10] C. Zhang, X. Zhang, D. Tu, and Y. Wang, "Small object detection using deep convolutional networks: Applied to garbage detection system," *Journal of Electronic Imaging*, vol. 30, no. 4, p. 043013, 2021.

[11] K. R. Rajani, A. Gaddam, and J. Gaddam, "IoT-based smart waste management using deep reinforcement learning," *Preprints*, 2022.

[12] A. Mishra, S. Ghosh, and D. P. Jena, "Internet of Things based waste management system for smart cities: A real time route optimization for waste collection vehicles," *International Journal of Scientific Research in Computer Science and Engineering*, vol. 6, no. 1, pp. 37–42, 2019.

[13] D. Abuga and N. S. Raghava, "Real-time smart garbage bin mechanism for solid waste management in smart cities," *Sustainable Cities and Society*, vol. 75, p. 103347, 2021.

[14] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal Policy Optimization Algorithms," *arXiv:1707.06347*, 2017.

[15] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, pp. 529–533, 2015.

[16] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. MIT Press, 2018.

[17] Y. Li, "Deep reinforcement learning: An overview," *arXiv preprint arXiv:1701.07274*, 2017.

[18] L. D. Xu, W. He, and S. Li, "Internet of Things in industries: A survey," *IEEE Transactions on Industrial Informatics*, vol. 10, no. 4, pp. 2233–2243, 2014.

[19] A. Martikkala, B. Mayanti, P. Helo, A. Lobov, and I. F. Ituarte, "Smart textile waste collection system—Dynamic route optimization with IoT," *Journal of Environmental Management*, vol. 335, p. 117548, 2023.

[20] S. R. J. Ramson, D. J. Moni, S. Vishnu, T. Anagnostopoulos, A. A. Kirubaraj, and X. Fan, "An IoT-based bin level monitoring system for solid waste management," *Journal of Material Cycles and Waste Management*, vol. 23, no. 2, pp. 516–525, 2021.

[21] D. Baldo, A. Mecocci, S. Parrino, G. Peruzzi, and A. Pozzebon, "A multi-layer LoRaWAN infrastructure for smart waste management," *Sensors*, vol. 21, no. 8, p. 2600, 2021.

[22] S. Karki, S. R. Niraula, and S. Karki, "Perceived risk and associated factors of healthcare waste in selected hospitals of Kathmandu, Nepal," *PLOS ONE*, vol. 15, no. 7, p. e0235982, 2020.
[23] H. Wu, F. Tao, and B. Yang, "Optimization of vehicle routing for waste collection and transportation," *International Journal of Environmental Research and Public Health*, vol. 17, no. 14, p. 4963, 2020.

LATIN-AMERICAN JOURNAL OF COMPUTING (LAJC), Vol XIII, Issue 2, July 2026

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.06

AUTHORS

Muhammad Masood Usmani

Muhammad Masood ul Rahman Usmani received his M.S. degree in Computer Science from COMSATS University Islamabad, Sahiwal Campus, Pakistan, in 2024, following his Master of Computer Science degree from The Islamia University of Bahawalpur, Pakistan. He is currently serving as a Faculty Member at Bahauddin Zakariya University, Lodhran Sub Campus, Pakistan, where he teaches undergraduate courses in Computer Science. His research interests include Machine Learning, Deep Learning, Reinforcement Learning, the Internet of Things (IoT), Computer Vision, and Artificial Intelligence for healthcare. He has authored and coauthored several research articles published in or submitted to international journals and conferences. His current research focuses on intelligent IoT systems, medical image analysis, and deep reinforcement learning based optimization for smart environments.

Rimsha Rafiq

Rimsha Rafiq received her M.S. degree in Computer Science from COMSATS University Islamabad, Sahiwal Campus, Pakistan, in 2024, following her Bachelor degree in Computer Science from Bahauddin Zakariya University, Multan, Pakistan. She is currently serving as a Teaching Faculty at COMSATS University Islamabad, Sahiwal Campus, Pakistan, where she teaches undergraduate courses in Computer Science. Her research interests include Machine Learning, Deep Learning, Reinforcement Learning and Artificial Intelligence for healthcare. Her current research focuses on medical image analysis and deep reinforcement learning based optimization for smart environments.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026

AUTHORS

Makki Riaz Khan

Makki Riaz Khan is a BS Information Technology student at Bahauddin Zakariya University, Sub Campus Lodhran, Pakistan. His research interests include Artificial Intelligence, Machine Learning, Deep Learning, and Computer Vision. He is a co-author of the paper "Multi-Class Classification of Alzheimer's Impairment using EfficientNet-B0". His technical expertise includes Python, TensorFlow, Keras, Scikit-learn, and OpenCV. He has developed several AI-based applications in healthcare and computer vision and is passionate about applying intelligent technologies to solve real-world problems.

M. M. U. R. Usmani, R. Rafiq, and M. R. Khan, "IoT-Enabled Deep Reinforcement Learning for Adaptive Waste Management in Hospital Environments", Latin-American Journal of Computing (LAJC), vol. 13, no. 2, 2026.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026

PAGE: Prompt Augmentation for Text Generation Enhancement

ARTICLE HISTORY
Received 13 October 2025
Accepted 5 January 2026
Published 7 July 2026

Mauro José Pacchiotti
Universidad Tecnológica Nacional
Centro de I+D de Ing. en Sistemas de Información
Santa Fe, Argentina
mpacchiotti@frsf.utn.edu.ar
ORCID: 0000-0002-9162-7890

Luciana Ballejos
Universidad Tecnológica Nacional
Centro de I+D de Ing. en Sistemas de Información
Santa Fe, Argentina
lballejos@frsf.utn.edu.ar
ORCID: 0000-0001-5443-6617

Mariel Ale
Universidad Tecnológica Nacional
Centro de I+D de Ing. en Sistemas de Información
Santa Fe, Argentina
male@frsf.utn.edu.ar
ORCID: 0000-0002-4866-4821

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026

M. Pacchiotti, L. Ballejos, and M. Ale
"PAGE: Prompt augmentation for text generation enhancement",
Latin-American Journal of Computing (LAJC), vol. 13, no. 2, 2026.

PAGE: Enriquecimiento de Prompts para mejorar la Generación de Texto

PAGE: Prompt Augmentation for Text Generation Enhancement

Mauro José Pacchiotti
Universidad Tecnológica Nacional
Centro de I+D de Ing. en Sistemas de Información
Santa Fe, Argentina
mpacchiotti@frsf.utn.edu.ar

Luciana Ballejos
Universidad Tecnológica Nacional
Centro de I+D de Ing. en Sistemas de Información
Santa Fe, Argentina
lballejos@frsf.utn.edu.ar

Mariel Ale
Universidad Tecnológica Nacional
Centro de I+D de Ing. en Sistemas de Información
Santa Fe, Argentina
male@frsf.utn.edu.ar

***Resumen*—** En los últimos años, los modelos generativos de lenguaje natural han demostrado un rendimiento sobresaliente en tareas de generación de texto. Sin embargo, cuando se enfrentan a tareas específicas o con requerimientos particulares, pueden presentar rendimientos pobres o necesitar ajustes que requieren grandes cantidades de datos adicionales. Este trabajo propone PAGE (Prompt Augmentation for text Generation Enhancement), un marco de trabajo que permite asistir a estos modelos mediante el uso de módulos auxiliares simples. Estos módulos, modelos simples como clasificadores o extractores, permiten obtener inferencias a partir del texto de entrada. La salida de estos auxiliares se utiliza para construir una entrada enriquecida que permite mejorar la calidad o controlabilidad de la generación. A diferencia de otras propuestas de asistencia a la generación, PAGE no exige el uso de modelos generativos auxiliares, sino que propone una arquitectura más simple, modular y fácil de adaptar a distintas tareas. Este artículo describe la propuesta, sus componentes y arquitectura, y presenta una prueba conceptual en el dominio de ingeniería de requerimientos, donde se utiliza un módulo auxiliar con un clasificador para mejorar la calidad en la generación de requerimientos de software.

***Palabras clave*—** *Generación de Requerimientos, LLM, Enriquecimiento de Prompts, PAGE*

***Abstract*—** In recent years, natural language generative models have shown outstanding performance in text generation tasks. However, when facing specific tasks or particular requirements, they may exhibit poor performance or require adjustments that demand large amounts of additional data. This work introduces PAGE (Prompt Augmentation for text Generation Enhancement), a framework designed to assist these models through the use of simple auxiliary modules. These modules—lightweight models such as classifiers or extractors—provide inferences from the input text. The output of these auxiliaries is then used to construct an enriched input that improves the quality and controllability of the generation. Unlike other generation-assistance approaches, PAGE does not require auxiliary generative models; instead, it proposes a simpler, modular architecture that is easy to adapt to different tasks. This paper presents the proposal, its components and architecture, and reports a proof of concept in the domain of requirements engineering, where an auxiliary module with a classifier is used to improve the quality of software requirements generation.

***Keywords*—** *Requirements Generation, LLM, Prompt Augmentation, PAGE*

I. Introducción

La generación de texto se ha convertido en una de las tareas más relevantes dentro del procesamiento de lenguaje natural, gracias a los avances en los grandes modelos lingüísticos (LLM, por sus siglas en inglés) como T5 (Text-to-Text Transfer Transformer) [1], GPT (Generative Pretraining Transformer) [2] o Llama [3] entre otros. Estos modelos, entrenados sobre grandes corpus, son capaces de generar texto con fluidez y coherencia. Sin embargo, cuando se aplican en tareas específicas donde se requiere que la salida cumpla con ciertas condiciones, limitaciones o estilos particulares, los resultados no siempre son favorables.

La problemática descripta -en algunos casos- puede solucionarse con el rentrenamiento o ajuste de parte o la totalidad del modelo. Aunque esta decisión de utilizar el entrenamiento de los modelos para mejorar el desempeño en una tarea tiene dos grandes implicancias. Por un lado, hace falta reunir y conformar conjuntos de datos con la cantidad y calidad suficiente de muestras, y por otro, se requiere del poder de cómputo necesario para la tarea de entrenamiento. Estas necesidades implican la disponibilidad de recursos que a veces están disponibles y, por lo tanto, no es posible lograr el objetivo propuesto.

Frente a las dificultades y necesidades descriptas, este trabajo propone PAGE, una arquitectura que incorpora módulos auxiliares que permiten obtener inferencias del texto de entrada. Estos módulos auxiliares pueden ser clasificadores, analizadores o extractores de características, entre otras opciones. Al ejecutarse antes del modelo generativo, los modelos auxiliares aportan metadatos o información estructurada que se incorpora a la entrada del generador. La arquitectura es modular y puede adaptarse según la tarea a resolver, lo que permite combinar distintos tipos de módulos auxiliares según las necesidades del dominio y la tarea a realizar.

El aporte de los módulos auxiliares pretende mejorar la salida del modelo generador y consumir menos recursos, tanto en el uso para generación, como también para el

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.07

entrenamiento, ya que pueden usarse auxiliares simples que no requieren de gran poder de cómputo o grandes conjuntos de datos de entrenamiento.

El resto del trabajo está organizado de la siguiente manera: la Sección II presenta el marco teórico y trabajos relacionados, la Sección III describe la herramienta PAGE, la Sección IV desarrolla una prueba conceptual centrada en la generación de requerimientos estructurados con sintaxis EARS [4]. Finalmente, en la Sección V se informan los resultados, mientras que en la Sección VI se exponen las conclusiones y se proponen posibles líneas de trabajos futuros.

II. Marco Teórico y Trabajos Relacionados

En tiempos recientes surgieron algunas propuestas para mejorar la entrada a un modelo generativo en busca de una mejor salida. Si bien varias propuestas utilizan modelos para asistir al generador, son variadas las formas de aplicación posibles. En esta línea, Du et al. [5] proponen mejorar LLMs en tareas de inferencia textual combinando prompts explícitos con conocimiento semántico extraído de bases externas. Utilizan atributos inferidos como insumos auxiliares, lo que mejora la precisión y coherencia de las respuestas. Por otro lado, He et al. [6] plantean asistir la generación de GPT-2 mediante resúmenes humanos codificados con BERT, guiando así la generación hacia mayor coherencia temática. Se evalúan distintas arquitecturas híbridas, con mejoras moderadas. En otro trabajo, Zeldes et al. [7] introducen Auxiliary Tuning, una técnica que adapta LLMs preentrenados a nuevas tareas, agregando un modelo auxiliar que ajusta la distribución de salida. La combinación se realiza a nivel de logits, sin modificar los pesos originales.

Más recientemente, Zhang et al. [8] proponen IAG (Induction-Augmented Generation), donde se utiliza un modelo generativo auxiliar para inducir conocimiento a partir del contexto, que luego se incorpora como entrada a otro modelo generativo. Esta inducción ha demostrado buenos resultados en tareas de razonamiento y QA. En otra propuesta, Liao et al. [9] plantean Awakening Augmented Generation, una técnica que activa el conocimiento latente en LLMs mediante tareas auxiliares previas, mejorando respuestas en QA. Su enfoque demuestra que pequeñas intervenciones bien diseñadas pueden guiar la generación sin alterar los parámetros base.

A diferencia de estos trabajos, PAGE propone utilizar módulos auxiliares simples y construir con sus salidas una entrada enriquecida que el modelo generativo pueda utilizar para mejorar sus respuestas. Esto permite una mayor interpretabilidad, modularidad y adaptabilidad, haciendo posible su implementación en escenarios con recursos limitados.

A. EARS

La propuesta de EARS [4] se basa en la identificación de patrones recurrentes en los requerimientos. A partir de un análisis empírico de especificaciones reales, los autores establecieron un conjunto reducido de plantillas sintácticas que cubren la mayoría de los casos prácticos. Estas plantillas permiten expresar diferentes categorías de requerimientos: Ubiquitous, Event-driven, State-driven, Optional y Unwanted, utilizando una sintaxis clara, que guía al analista en la redacción de cada expresión. De este modo, se logra un lenguaje controlado que reduce la ambigüedad, sin exigir conocimientos técnicos avanzados en lenguajes formales.

Uno de los beneficios principales de EARS es que facilita la comunicación entre stakeholders. Al proporcionar una estructura reconocible y repetible, se reduce la presencia de ambigüedad y se mejora la trazabilidad de los requerimientos a lo largo del ciclo de vida del software. Asimismo, la simplicidad de la técnica permite que usuarios no especializados participen en la redacción y revisión de las especificaciones.

B. Grandes Modelos Lingüísticos

A partir del trabajo *Attention is all you need* [10] surge el modelo Transformer, un enfoque de aprendizaje profundo cuya arquitectura se organiza en dos estructuras principales: un codificador y un decodificador, ambos basados en el mecanismo de atención multicabeza. A diferencia de los modelos recurrentes que procesan el texto palabra por palabra en orden secuencial, el Transformer puede considerar todas las palabras de una oración al mismo tiempo. Gracias a este mecanismo, los modelos lingüísticos pueden resaltar las partes más relevantes de una entrada y comprender mejor tanto su significado como su contexto. Esto hace posible capturar relaciones de largo alcance entre términos, lo cual se traduce en mejoras sustanciales en múltiples tareas de procesamiento del lenguaje natural.

A partir del Transformer se desarrollaron distintos modelos que adoptaron y expandieron esta arquitectura, como BERT (Bidirectional Encoder Representations from Transformers) [11], T5 [1], GPT [2] y Llama [3] entre otros. Estos modelos se entrenaron con volúmenes cada vez mayores de datos, incluyendo texto y código, y se destacaron por su capacidad para generar texto de alta calidad y adaptarse a un amplio rango de tareas en PLN. La evolución continuó con la aparición de ChatGPT [12] en 2022. Esta aplicación de OpenAI llevó el uso de la IA generativa a un público amplio a través de una interfaz de chat, basada en la familia de modelos GPT, capaz de generar respuestas coherentes y contextuales en lenguaje natural.

C. Técnicas de Prompting

De esta interacción con LLMs, surge la posibilidad de comunicarse con un modelo mediante expresiones en lenguaje natural. En este contexto, la entrada que el usuario proporciona recibe el nombre de prompt, entendido como una instrucción o conjunto de palabras que orientan la generación de la respuesta. La manera en que se formula este prompt resulta fundamental, ya que condiciona directamente la salida producida por el modelo. Por esto, distintas guías de buenas prácticas señalan qué elementos conviene considerar al redactarlo con el fin de obtener resultados más cercanos a lo esperado (ver Tabla I).

TABLA I. Elementos recomendados en un prompt [13]

Elemento	Descripción
Instrucción	Tarea específica que se desea.
Contexto	Información adicional que puede orientar al modelo y completar la respuesta.
Entrada	La entrada sobre la que se desea la acción.
Salida	Formato que se desea para la respuesta del modelo.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.07

Es importante destacar que estos elementos no son estrictamente obligatorios, sino que su inclusión depende de la necesidad en cada caso. Asimismo, el prompt puede enriquecerse incorporando ejemplos de la tarea deseada, lo que permite al modelo imitar de manera más precisa el comportamiento esperado. De acuerdo con la cantidad de ejemplos aportados, se reconocen tres enfoques principales: zero-shot (sin ejemplos), one-shot (con un ejemplo) y few-shot (con varios ejemplos). Entre ellos, el enfoque few-shot suele ser el más potente, ya que mejora la capacidad de generalización del modelo y produce salidas de mayor calidad.

D. ROUGE

La evaluación automática de sistemas de generación de texto requiere métricas que permitan medir la calidad de una salida en comparación con referencias humanas. En este ámbito, ROUGE (Recall-Oriented Understudy for Gisting Evaluation), propuesta por Lin [14], se consolidó como una de las técnicas más utilizadas en el análisis de resúmenes automáticos y se extiende a la evaluación de modelos generativos. La métrica se basa en la superposición de unidades de texto (n-gramas, secuencias o subsecuencias) entre el texto generado y uno o más textos de referencia, midiendo así la similitud entre ellos.

Entre las variantes más empleadas se encuentran ROUGE-1, ROUGE-2 y ROUGE-L. ROUGE-1 evalúa la coincidencia de unigramas (palabras individuales) entre el texto generado y la referencia, proporcionando una medida básica de cobertura del contenido. ROUGE-2, por su parte, se centra en la coincidencia de bigramas, lo que introduce un nivel mayor de sensibilidad al orden y la fluidez de las palabras, capturando relaciones locales entre términos. Finalmente, ROUGE-L se basa en la subsecuencia más larga de palabras en común entre las cadenas comparadas (LCS en inglés: Longest Common Subsequence), lo que permite valorar la preservación de la estructura y el orden global de la información [14].

Un aspecto importante de ROUGE es que puede calcularse en términos de recall (1), precisión (2) y F1-score (3), aunque en el ámbito de generación de texto se utiliza más frecuentemente el recall, al priorizar la recuperación del contenido presente en el texto de referencia.

$$\boldsymbol{Recall_{ROUGE}} = \frac{\boldsymbol{n-gramas\ coincidentes}}{\boldsymbol{n-gramas\ en\ la\ referencia}} \tag{1}$$

$$\boldsymbol{Precision_{ROUGE}} = \frac{\boldsymbol{n-gramas\ coincidentes}}{\boldsymbol{n-gramas\ en\ la\ generación}} \tag{2}$$

$$\boldsymbol{F1Score_{ROUGE}} = \frac{\boldsymbol{2.Precicision.Recall}}{\boldsymbol{Precision+Recall}} \tag{3}$$

III. HERRAMIENTA PROPUESTA

La propuesta PAGE parte de la idea de que los modelos generativos pueden beneficiarse al recibir como entrada no sólo el texto original, sino también información extra estructurada inferida del mismo texto. Estas inferencias pueden obtenerse mediante módulos auxiliares específicos, modelos o algoritmos, diseñados según la tarea de generación a resolver. Estos módulos auxiliares permiten, a partir de sus inferencias mejorar la salida del modelo generador. Si se analiza desde la perspectiva de los recursos, se traduce en una mejora en la tarea de generación sin entrenar o ajustar un LLM, sino pequeños modelos auxiliares. Se trata de una alternativa que pretende mejorar la generación de un modelo reduciendo la necesidad de grandes conjuntos de datos y capacidades de cómputo para realizar ajustes o entrenamientos.

PAGE se estructura como una arquitectura modular compuesta por tres componentes principales: el conjunto de módulos auxiliares, el compositor de contexto, y el generador principal. Cada componente puede adaptarse según la aplicación concreta, permitiendo utilizar módulos simples y muy interpretables como clasificadores, etiquetadores, extractores o analizadores de sentimientos, entre otras opciones. También pueden ser útiles funciones que aporten de acuerdo con la expresión original, información de alguna fuente externa.

A. Módulo Auxiliar

El módulo auxiliar es el componente responsable de realizar una o más inferencias sobre el texto de entrada, con el fin de obtener información estructurada que luego será utilizada para asistir a la generación. A diferencia del modelo generativo principal, estos modelos son generalmente más simples, entrenados para tareas específicas y diseñados para ser fácilmente interpretables. Además, al ser modelos más sencillos requieren un menor esfuerzo de entrenamiento, tanto desde el punto de vista del poder de cómputo como de los datos. El tipo de modelo auxiliar a emplear depende directamente del dominio y los objetivos de generación. A continuación, se describen algunos tipos comunes:

- *Clasificador:* Un modelo que clasifica con alguna etiqueta de interés para la tarea el texto de entrada. Existe una gran variedad de modelos y pipelines que pueden realizar esta tarea, utilizando desde simples modelos estadísticos hasta redes neuronales profundas o grandes modelos lingüísticos.

- *Extractor de entidades o partes del discurso:* Este tipo de modelo identifica y clasifica entidades, acciones o porciones relevantes dentro del texto. En estos casos la salida podría tratarse de una estructura con los tokens extraídos.

- *Analizador de sentimiento o intención:* Se trata de un tipo de clasificador que permite detectar el tono, la urgencia o la finalidad de una necesidad, así como también el sentimiento que expresa el autor del texto. Existen varias propuestas para las etiquetas de salida en estos casos.

En todos los casos, la salida del módulo auxiliar debe ser expresada de forma explícita y legible, generalmente como texto estructurado, de modo que pueda ser utilizada sin preprocesamientos por el Compositor de Prompts. Esta estrategia facilita la trazabilidad del sistema y mantiene interfaces claras, lo que resulta importante para las actividades de gestión y control del proceso.

B. Compositor de Prompts

Este artefacto toma la salida de los módulos auxiliares, que pueden ser uno o más, y construye un bloque de entrada mejorado que se combina con el texto original. Esta composición puede seguir plantillas configurables o estructuras semiformales según el objetivo de la instrucción y las estructuras devueltas por los módulos auxiliares.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.07

Finalmente, la salida es el texto que se utiliza como entrada en el modelo generativo.

C. Modelo Generativo

El componente generativo puede ser cualquier LLM capaz de producir texto a partir de la entrada de un prompt. Si bien puede efectuarse un ajuste fino o reentrenamiento del modelo generativo, este no siempre resulta necesario; se espera que el enriquecimiento contextual, correctamente estructurado, sea suficiente para dirigir la generación hacia una salida con la calidad deseada.

Dependiendo de la aplicación pueden utilizarse distintas técnicas de prompting para utilizar el modelo generativo, dentro de las que se destacan las que tienen que ver con los ejemplos que se proveen al modelo, One-shot y Few-shots.

D. Proceso

El flujo de trabajo en PAGE se resume en los siguientes pasos:

1) Un usuario o sistema proporciona un texto base.
2) Los módulos auxiliares procesan esta entrada y generan información estructurada a partir de sus inferencias.
3) El compositor integra la información estructurada con el texto original y construye un prompt enriquecido mediante una plantilla.
4) El modelo generativo produce la salida final a partir de esta entrada enriquecida.

El flujo del proceso (Fig. 1) permite realizar pruebas de los componentes por separado para analizar el impacto de cada auxiliar, así como adaptar el marco a distintos dominios sin modificar el modelo generativo. Además, la naturaleza textual de los componentes auxiliares facilita la depuración, interpretación y validación de los pasos intermedios.

IV. Prueba Conceptual

Para probar la propuesta se diseñó una implementación del marco con el objetivo de mejorar expresiones de requerimientos de software. Son diversas las propuestas sobre la estructura sintáctica que se debe utilizar para expresar requerimientos, con el objetivo de reducir la ambigüedad y otras deficiencias. El enfoque EARS [4] ofrece clasificar los requerimientos de software en cinco categorías: Event-driven, Ubiquitous, State-driven, Unwanted behavior y Optional; para posteriormente emplear una plantilla particular para cada tipo, lo que permite guiar y restringir la generación de las expresiones.

A. Conjunto de Datos

Las pruebas se realizan sobre un Dataset que resulta de la recopilación de textos de requerimientos desde diversas fuentes, los Datasets *PURE* [15] y *Software Functional Requirements* [16], así como también requerimientos obtenidos desde diversos documentos de especificación de dominio público. Cabe resaltar que se buscó diversidad de dominios y balance con respecto a las categorías de la propuesta EARS [4] (Figura 2).

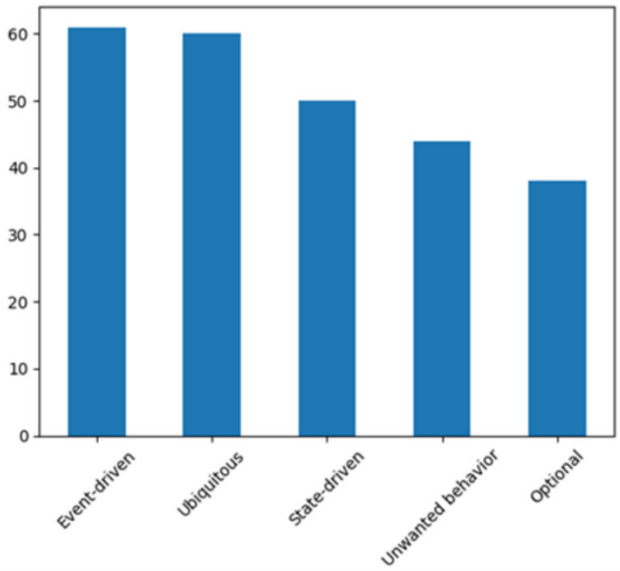


Fig. 2. Balance del conjunto de datos

La estructura del dataset está compuesta por tres columnas: a) la expresión de requerimiento sin una estructura sintáctica definida, b) la etiqueta correspondiente con la categoría de EARS y c) la expresión con sintaxis EARS elaborada manualmente, conforme a las indicaciones del enfoque. El conjunto de datos empleado consta de 253 instancias, un tamaño reducido que resulta apropiado para evaluar en qué medida la propuesta permite disminuir el esfuerzo asociado a la preparación de conjuntos de datos para entrenamiento.

B. Componentes

En esta implementación de PAGE, con el fin de generar expresiones de requerimientos según la propuesta EARS, se definen los siguientes componentes:

Modulo auxiliar: Para esta implementación se utiliza un solo módulo auxiliar que contiene un clasificador. Este modelo simple, dada una expresión textual, devuelve la etiqueta correspondiente a la categoría EARS. Luego, la etiqueta es utilizada para que el módulo devuelva ejemplos correspondientes con esa categoría que puedan ser anexados como información de contexto al prompt.

Para disponer de un modelo clasificador que pueda entrenarse con pocas muestras, se realizó un entrenamiento

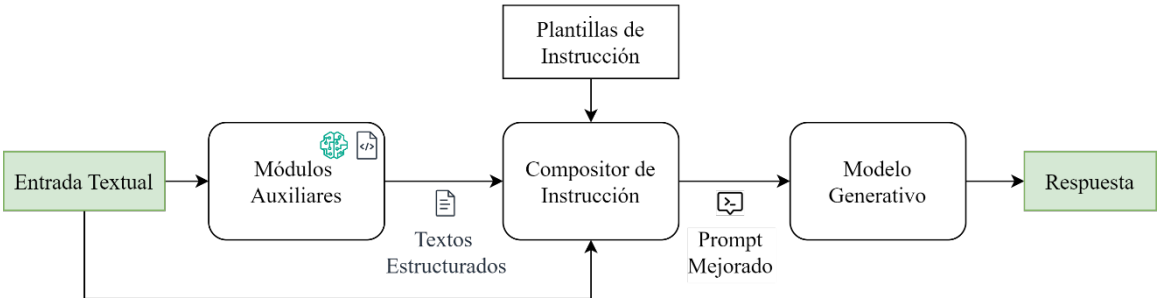


Fig. 1. Proceso de PAGE

con búsqueda de grilla para ajuste de hiperparámetros de un modelo Random Forest [17]. Este tipo de modelos es una de las técnicas de aprendizaje supervisado más utilizadas frente a conjuntos de datos reducidos. Al basarse en un ensamble de árboles de decisión entrenados sobre subconjuntos de datos y características, logra disminuir el riesgo de sobreajuste que suelen presentar los modelos individuales. Esta propiedad lo convierte en una alternativa confiable cuando se dispone de un número limitado de muestras, ya que aprovecha la variabilidad introducida por el muestreo y mantiene un equilibrio entre sesgo y varianza [17]. Además, cumple con una de las motivaciones de la propuesta, por tratarse de un modelo que requiere de muy poco poder de cómputo para su entrenamiento.

La configuración de hiperparámetros que obtuvo el mejor desempeño para el modelo Random Forest correspondió a una profundidad máxima de 10, un mínimo de 5 muestras por división y 100 estimadores. Para el entrenamiento, el conjunto de datos se particionó reservando un 20% para pruebas, complementado con un esquema de validación cruzada de cinco particiones. Con esta configuración, el modelo alcanzó un accuracy del 82.35% sobre el conjunto de test. La Figura 3 muestra las medidas de performance obtenidas y la Figura 4 la matriz de confusión.

	precision	recall	f1-score	support
Event-driven	0.83	0.83	0.83	12
Optional	0.88	0.88	0.88	8
State-driven	1.00	0.80	0.89	10
Ubiquitous	0.69	0.92	0.79	12
Unwanted behavior	0.86	0.67	0.75	9
accuracy			0.82	51
macro avg	0.85	0.82	0.83	51
weighted avg	0.84	0.82	0.82	51

Fig. 3. Resultados del modelo de clasificación con el conjunto de Test

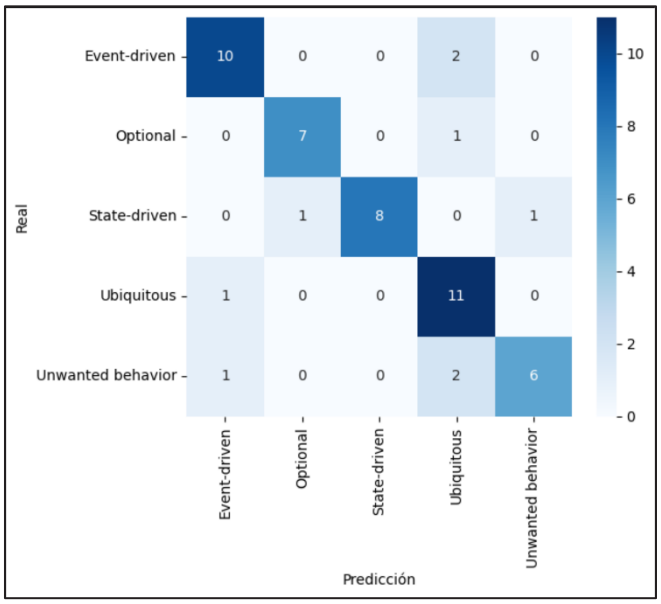


Fig. 4. Matriz de confusión de los resultados con el conjunto de Test

Esta etiqueta obtenida por el clasificador se utiliza como parámetro para que el módulo devuelva dos ejemplos de requerimientos escritos según la sintaxis EARS para esa categoría. La Tabla II muestra los ejemplos utilizados en las pruebas, de acuerdo con la categoría EARS.

Compositor de contexto: El propósito de este componente es generar el prompt mejorado que será ingresado al modelo generativo. Para este ejemplo se utiliza una plantilla (Figura 5) que permite anexar los ejemplos devueltos por el módulo auxiliar en la etiqueta {examples_text}.

TABLA I I. EJEMPLOS SEGÚN LA CATEGORÍA EARS

Categoría	Ejemplos
Ubiquitous	**requirement:** "The system shall log all transactions." **ears:** "The system shall always log all transactions." **requirement:** "The application shall keep the session active during user activity.", **ears:** "The system shall always keep the session active during user activity."
Event-driven	**requirement:** "The system shall notify the admin when the server restarts.", **ears:** "When the server restarts, the system shall notify the admin." **requirement:** "The application shall send a receipt when a purchase is completed.", **ears:** "When a purchase is completed, the application shall send a receipt."
State-driven	**requirement:** "The system shall block new logins while maintenance mode is active." **ears:** "While maintenance mode is active, the system shall block new logins." **requirement:** "The application shall allow offline access while the device has no internet connection." **ears:** "While the device has no internet connection, the application shall allow offline access."
Unwanted behavior	**requirement:** "The system shall display a warning if unauthorized access is detected.", **ears:** "If unauthorized access is detected, the system shall display a warning.". **requirement:** "The application shall stop the upload if the file exceeds the maximum size.", **ears:** "If the file exceeds the maximum size, the application shall stop the upload."
Optional	**requirement:** "The system shall enable voice control where the device supports it.", **ears:** "Where the device supports it, the system shall enable voice control." **requirement:** "The application shall provide dark mode where the user has selected the option.", **ears:** "Where the user has selected the option, the application shall provide dark mode."

Modelo generativo: En esta implementación, se utiliza un modelo generativo con licencia de uso público, Llama 3.1[1] en su versión con 8 billones de parámetros entrenables. Para implementarlo se despliega sobre la herramienta Ollama[2] que permite ejecutarlo de manera local y consumirlo como un servicio desde un entorno Jupyter Notebook[3] con el lenguaje de programación Python[4].

Experimentos: Se realizaron tres pruebas sobre el dataset completo, compuesto por 253 filas, con el objetivo de recolectar resultados que permitan validar la utilidad de la propuesta. La primera prueba consistió en utilizar únicamente

[1] https://huggingface.co/meta-llama/Llama-3.1-8B
[2] https://ollama.com/
[3] https://jupyter.org/
[4] https://www.python.org/

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.07

```
You are an assistant that rewrites requirements using the EARS syntax.
Rewrite the following requirement using the EARS syntax.
Use the examples below as a guide.
{examples_text}
--------------------------------------------------------------------------
Respond ONLY with the rewritten requirement. Do not add explanations, comments, or any extra text.Requirement:

{natural}

EARS Requirement:
```

Fig. 5. Plantilla para generación del prompt

el modelo generativo junto con la expresión textual original y un prompt few-shot fijo, diseñado para cubrir explícitamente todas las categorías de la sintaxis EARS mediante ejemplos representativos, sin emplear ningún mecanismo auxiliar de inferencia. Esta configuración permitió establecer una línea base fuerte, en la que el modelo debe inferir la estructura adecuada para la generar la expresión a partir del contexto provisto por el prompt. En la segunda, se incorporó el Compositor de Prompts y una versión ideal del Módulo Auxiliar, que disponía de la etiqueta correcta para cada expresión proveniente del dataset. Este diseño permitió obtener simultáneamente el piso de desempeño (prompt few-shot fijo) y el techo alcanzable (con enriquecimiento derivado de las etiquetas correctas provistas por el dataset). Finalmente, la tercera prueba implementó el proceso PAGE completo, utilizando el Módulo Auxiliar descripto en la Sección 4.2.

V. Resultados

Con el objetivo de analizar el comportamiento de la propuesta, se realizaron tres instancias de evaluación. En primer lugar, se calcularon métricas automáticas basadas en ROUGE, para cuantificar el grado de solapamiento léxico entre las expresiones generadas y las expresiones correctas provistas por el dataset. En segundo lugar, se ejecutó un procedimiento de validación automática para evaluar el cumplimiento de restricciones de la sintaxis EARS en las estructuras de las expresiones generadas. Finalmente, se realizó una evaluación manual por parte de un experto sobre una muestra de 30 expresiones, considerando para cada una de ellas la salida producida por los tres enfoques evaluados. El objetivo fue complementar las métricas automáticas y obtener una apreciación cualitativa respecto de la completitud, la ambigüedad y el cumplimiento de la plantilla EARS.

Para las pruebas iniciales se utilizaron como métricas ROUGE 1, ROUGE 2 y ROUGE L calculando para cada una el recall, la precisión y el F1-Score, comparando cada expresión con la correcta. La Tabla III muestra los resultados obtenidos en cada prueba.

TABLA III. Resultados para el modelo sin módulos auxiliares (Few-Shot) con módulo auxiliar basado en la etiqueta del dataset (Dataset-samples) y para la propuesta PAGE (PAGE)

Experimento	Métrica	Precisión	Recall	F1-Score
Few-Shot	ROUGE1	0,813	0,821	0,807
Few-Shot	ROUGE2	0,636	0,643	0,632
Few-Shot	ROUGEL	0,747	0,749	0,739
Dataset-samples	ROUGE1	0,852	0,815	0,827
Dataset-samples	ROUGE2	0,653	0,630	0,636
Dataset-samples	ROUGEL	0,803	0,770	0,781
PAGE	ROUGE1	0,849	0,809	0,822
PAGE	ROUGE2	0,648	0,622	0,630
PAGE	ROUGEL	0,796	0,761	0,772

En los resultados, se destacan los valores elevados obtenidos con la métrica ROUGE-1, lo que sugiere que los modelos capturan adecuadamente los términos individuales presentes en los requerimientos. Asimismo, los valores de ROUGE-2 y ROUGE-L indican que también logran reproducir combinaciones locales de palabras y estructuras más largas. En este contexto, PAGE alcanza un desempeño comparable al techo alcanzable, manteniendo niveles

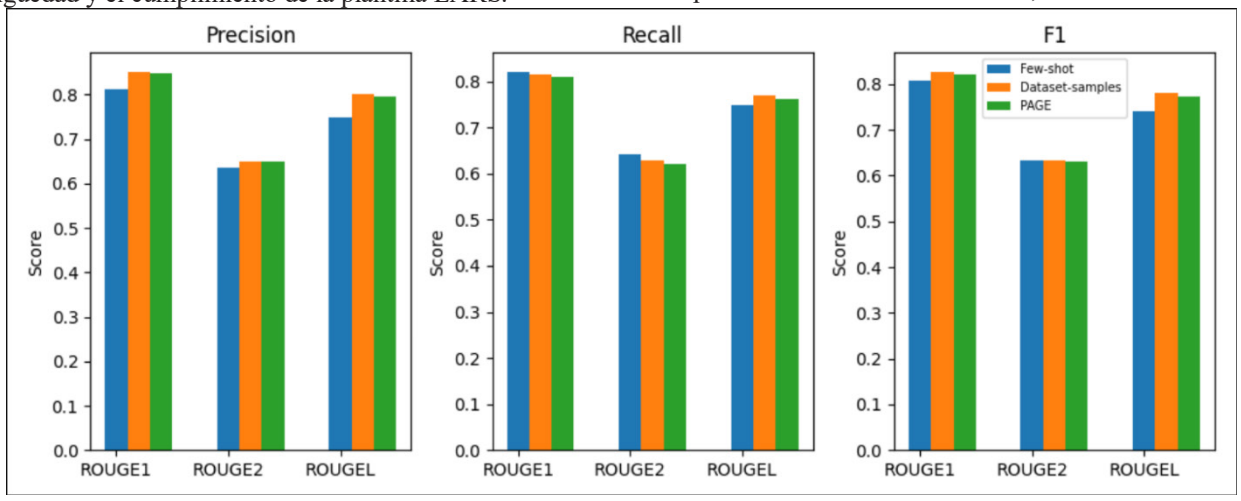


Fig. 6. Rendimientos obtenidos en las tres pruebas realizadas

similares tanto en términos individuales como en la captura de combinaciones locales de palabras. La Figura 6 presenta de manera gráfica las diferencias observadas entre los distintos enfoques evaluados.

Para la segunda prueba se incorporó una validación automática orientada a evaluar en que proporción las expresiones generadas cumplen las estructuras propuestas en la sintaxis EARS. Este procedimiento adopta un enfoque determinístico, y se limita a verificar el orden relativo de marcadores textuales claves definidos por las plantillas EARS, sin evaluar corrección semántica o roles gramaticales.

La validación se realizó teniendo como condición la categoría EARS a la que corresponde la expresión, mediante la aplicación de un conjunto de reglas y utilizando expresiones regulares. Para la categoría Ubiquitous, se comprobó que el requerimiento siga el patrón básico "The <texto> shall <texto>". Para el resto de las categorías, se verificó que el texto comience con la palabra clave correspondiente y que esta anteceda al verbo "shall". La tabla IV muestra los patrones verificados en esta prueba.

TABLA IV. ESTRUCTRURAS EVALUADAS EN LA PRUEBA

Categoría EARS	Estructura verificada
Ubiquitous	The <texto> shall <texto>
Event-driven	WHEN <texto> shall <texto>
State-driven	IF <texto> shall <texto>
Unwanted behavior	WHILE <texto> shall <texto>
Optional	WHERE <texto> shall <texto>

Cada requerimiento se clasificó como conforme o no conforme, y la métrica se calculó como la proporción de salidas que respetaron la estructura esperada sobre el total evaluado.

Los resultados de esta prueba (Tabla V) muestran diferencias claras entre los enfoques evaluados. Las expresiones generadas con PAGE alcanzan un nivel de cumplimiento del 87%, superando ampliamente a la línea base "Few Shot" y acercándose al desempeño del techo alcanzable "Dataset-label", lo que sugiere que la selección dirigida de ejemplos con el aporte del módulo auxiliar permite mejorar significativamente la generación de expresiones que respetan la estructura.

TABLA V. RESULTADOS DE LA SEGUNDA PRUEBA

Experimento	Resultado (%)
Few-shot	58,5
Dataset-label	92,9
PAGE	87,0

Finalmente, se realizó una evaluación manual sobre una partición estratificada de 30 expresiones, considerando para cada una de ellas la salida producida por los tres enfoques evaluados y seleccionadas de forma aleatoria, cuidando mantener seis expresiones por cada categoría EARS. Si bien se trató de una prueba conceptual orientada a analizar cómo PAGE mejora la redacción de requerimientos según EARS, la evaluación manual no se limitó únicamente a verificar el cumplimiento de la plantilla EARS. Adicionalmente, se incorporaron como criterios la completitud y la no ambigüedad de las expresiones generadas, permitiendo así una valoración más integral de su calidad. Este enfoque complementa las métricas automáticas previamente presentadas, permitiendo una valoración más cualitativa de las expresiones generadas en cada prueba.

En las tres características evaluadas se puede observar un rendimiento menor de la propuesta con un prompt few-shot fijo (Tabla VI), esto se debe principalmente a que, aunque tenga el prompt ejemplos de todas las categorías, a veces, el modelo no puede identificar la plantilla correcta y fuerza la generación a una estructura que no corresponde, redundando u omitiendo información importante.

TABLA VI. RESULTADOS DE LA EVALUACIÓN MANUAL

Experimento	Completitud	No ambiguedad	Estructura EARS
Few-shot	53,3	60,0	66,7
Dataset-label	76,7	73,3	86,7
PAGE	76,7	73,3	86,7

Los resultados de PAGE y del modelo con etiquetas provistas por el dataset no presentan diferencias, las expresiones generadas eran idénticas y alcanzaron ambos un buen rendimiento. Esto probablemente se deba a que, en los ejemplos considerados en la partición seleccionada para la prueba, el modelo clasificador de la propuesta PAGE clasificó la etiqueta correcta, por lo que la salida de los modelos no difiere al utilizarse el mismo prompt, con los mismos ejemplos en ambos casos.

En términos generales, esta prueba conceptual se realizó sobre el mismo modelo y con el mismo conjunto de datos en todos los experimentos buscando evaluar el aporte del módulo auxiliar, pieza central de la propuesta PAGE. De este análisis también puede concluirse que el módulo auxiliar tiene gran influencia sobre la generación, así, cuando la clasificación falla, conduce al modelo generativo a una estructura sintáctica equivocada para esa clase de requerimiento. Esto demuestra que, en las implementaciones del marco, es importante dedicar esfuerzos a mejorar las respuestas de los módulos auxiliares para lograr aportes correctos al modelo generativo.

A. Limitaciones

La evaluación presentada en este trabajo está desarrollada como una prueba conceptual, por lo que presenta una serie de limitaciones que deben ser consideradas. En primer lugar, se realiza utilizando un único modelo generativo y un único módulo auxiliar. En segundo lugar, los experimentos se llevan a cabo sobre un conjunto de datos acotado y específico del dominio de los requerimientos de software, que resulta adecuado para ilustrar la viabilidad de la propuesta, pero no permite realizar afirmaciones sólidas respecto de su escalabilidad o robustez en otros contextos.

Asimismo, la evaluación se centra en aislar el impacto de la presencia del módulo auxiliar como principal variable, sin explorar de manera exhaustiva otras configuraciones posibles ni métricas de eficiencia como latencia o sobrecarga computacional. Estos aspectos, junto con evaluaciones humanas más extensas y experimentos con múltiples modelos y conjuntos de datos, quedan planteados como líneas de trabajo futuro. A pesar de estas limitaciones, los resultados obtenidos aportan evidencia que respalda la viabilidad del

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.07

marco PAGE como una estrategia para mejorar la generación de texto.

VI. Conclusiones y Trabajos Futuros

PAGE propone un enfoque sencillo y eficaz para mejorar el rendimiento, la controlabilidad y la estructura de las salidas generadas por LLMs, aprovechando inferencias y aportes simples y adaptables al contexto de uso. Los resultados obtenidos en las pruebas iniciales son prometedores y abren la posibilidad de replicar los experimentos con otros datasets y en diferentes dominios. Respecto de la prueba conceptual, se concluye que la solución planteada logra medidas alentadoras para las métricas evaluadas, lo que sugiere la conveniencia de ampliar la validación hacia conjuntos de datos con más requerimientos y en diferentes dominios.

El principal aporte de esta propuesta radica en demostrar que la generación de texto puede manipularse de manera efectiva mediante herramientas sencillas, apoyadas en módulos auxiliares simples y altamente interpretables. Estos módulos desempeñan un papel clave dentro del marco: pueden guiar al modelo hacia una salida adecuada o, en caso contrario, conducirlo a resultados erróneos, lo que resalta su relevancia y necesidad de un cuidadoso diseño.

Como línea de trabajo futuro, se plantea la especificación e implementación de un marco de software en Python que proporcione una estructura completamente reutilizable. Dicho marco deberá facilitar la implementación, evaluación y persistencia de distintas instancias de PAGE, facilitando así su aplicación práctica y la extensión de sus capacidades en nuevos contextos.

References

[1] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, arXiv preprint arXiv:1910.10683, 2019.

[2] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, and D. Amodei, Language Models are Few-Shot Learners, arXiv preprint arXiv:2005.14165, 2020.

[3] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, LLaMA: Open and Efficient Foundation Language Models, arXiv preprint arXiv:2302.13971, 2023.

[4] A. Mavin, P. Wilkinson, A. Harwood, and M. Novak, "Easy Approach to Requirements Syntax (EARS)," in Proc. 17th IEEE Int. Requirements Engineering Conf. (RE'09), 2009, pp. 317–322.

[5] L. Du, X. Ding, K. Xiong, T. Liu, and B. Qin, "Enhancing pretrained language models with structured commonsense knowledge for textual inference," Knowledge-Based Systems, vol. 254, p. 109488, 2022, doi: 10.1016/j.knosys.2022.109488.

[6] K. He, J. K. Chen, and P. Kim, "Generative pre-training language models with auxiliary conditional summaries," Stanford University, Stanford, CA, USA, Tech. Rep., 2020. [Online]. Available: https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1204/reports/custom/report50.pdf

[7] Y. Zeldes, D. Padnos, O. Sharir, and B. Peleg, "Auxiliary tuning and its application to conditional text generation," arXiv:2006.16823, 2020.

[8] Z. Zhang, X. Zhang, Y. Ren, S. Shi, M. Han, Y. Wu, R. Lai, and Z. Cao, "IAG: Induction-Augmented Generation Framework for Answering Reasoning Questions," in Proc. EMNLP 2023, 2023, pp. 1–14, doi: 10.18653/v1/2023.emnlp-main.1.

[9] H. Liao, S. He, Y. Xu, Y. Zhang, S. Liu, K. Liu, and J. Zhao, "Awakening Augmented Generation: Learning to awaken internal knowledge of large language models for question answering," in Proc. 31st Int. Conf. on Computational Linguistics (COLING), Abu Dhabi, UAE, 2025, pp. 1333–1352, Assoc. for Computational Linguistics.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding," in Proc. 2019 Conf. of the North American Chapter of the Assoc. for Computational Linguistics: Human Language Technologies (NAACL-HLT), 2019, vol. 1, pp. 4171–4186.

[12] OpenAI, "ChatGPT," 2025. [Online]. Available: https://chatgpt.com

[13] Prompt Engineering Guide, "Elements of a prompt," 2024. [Online]. Available: https://www.promptingguide.ai/introduction/elements

[14] C. Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in Proc. ACL Workshop on Text Summarization Branches Out, Barcelona, Spain, 2004, pp. 74–81.

[15] A. Ferrari, G. O. Spagnolo, and S. Gnesi, "PURE: A dataset of public requirements documents (version 2.0)," in Proc. 25th IEEE Int. Requirements Engineering Conf. (RE), Lisbon, Portugal, 2017. [Online]. Available: https://doi.org/10.5281/zenodo.7118517

[16] T. Tahir and K. Tasleem, "Software functional requirements," Zenodo, Dataset, 2025. [Online]. Available: https://doi.org/10.5281/zenodo.15834954

[17] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, 2001, pp. 5–32, doi: 10.1023/A:1010933404324.

AUTHORS

Mauro José Pacchiotti

Mauro José Pacchiotti es Ingeniero en Sistemas de Información y Especialista en Ingeniería en Sistemas de Información. Actualmente es alumno de posgrado del Doctorado en Ingeniería, Mención Sistemas de Información, en la Facultad Regional Santa Fe de la Universidad Tecnológica Nacional (UTN), República Argentina. Se desempeña como docente en la carrera de Ingeniería en Sistemas de Información de la UTN Facultad Regional Santa Fe y como docente-investigador en el Centro de Investigación y Desarrollo de Ingeniería en Sistemas de Información (CIDISI).

Sus principales áreas de interés comprenden la Inteligencia Artificial, la Ingeniería de Requerimientos y la Ciencia de Datos. Participa activamente en proyectos de investigación relacionados con estas áreas.

Es autor y coautor de diversos artículos científicos publicados en revistas especializadas y en actas de congresos. Además, ha participado en proyectos de transferencia de conocimiento y en actividades de extensión orientadas al medio socio-productivo. Es miembro estudiante del IEEE (Institute of Electrical and Electronics Engineers).

Luciana Ballejos

Ingeniera en Sistemas y Doctora en Ingeniería, mención Sistemas de Información. Es Profesora Titular Ordinaria y Directora de la carrera de Ingeniería en Sistemas de Información de la Facultad Regional Santa Fe (Universidad Tecnológica Nacional) y docente investigadora en el Centro de Investigación y Desarrollo de Ingeniería en Sistemas de Información (CIDISI). Es docente de cursos de posgrado y dirige e integra proyectos de investigación en el área de Ingeniería de Software, Aprendizaje Automático e Inteligencia Artificial, además de dirigir y codirigir tesis de posgrado en el área. Es par evaluadora de CONEAU de carreras de grado en el área de Informática, integrante de Comités Técnicos y de Programa de reuniones científicas nacionales e internacionales en el área y evaluadora externa de Comités de Evaluación Científica y Tecnológica en diversas universidades del país, además de evaluadora de trabajos en revistas internacionales. Integra equipos de trabajo que generan transferencia, extensión y servicios a terceros desde la universidad al medio y la región.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026

AUTHORS

Mariel Ale

Ingeniera en Sistemas de Información y Doctora en Ingeniería Mención Sistemas de Información de la UTN – FRSF, tiene categoría II en el Programa de Incentivos para docentes investigadores de la República Argentina. Actualmente es Profesora Titular Ordinaria de Ingeniería en Sistemas de Información y en varios cursos de posgrado. Además, se desempeña como Investigadora en el Centro de Investigación y Desarrollo de Ingeniería en Sistemas de Información (CIDISI) de dicha facultad. Dirige proyectos de investigación en el área de Ciencia de Datos, Aprendizaje Automático e Inteligencia Artificial y es autora de numerosos artículos publicados en revistas y actas de congresos. Tiene a su cargo becarios doctorales, de maestría y de grado. Es miembro de comités científicos de congresos nacionales e internacionales y revistas científicas de publicación periódica. Se desempeña como consejera departamental docente en la UTN – FRSF. Ha participado en diversos proyectos de transferencias de conocimiento y extensión al medio socio-productivo.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026

Enhancing Cybersecurity with Random Forest: Efficient Detection of Cyberattacks

ARTICLE HISTORY
Received 26 February 2026
Accepted 3 June 2026
Published 7 July 2026

Phathutshedzo Cyprin Ramuhovhi
Vaal University of Technology
Computer Sciences and Engineering Vanderbijlpark, South Africa
214102653@edu.vut.ac.za
ORCID: 0000-0001-9232-8852

Naume Sonhera
Vaal University of Technology
Computer Sciences and Engineering
Vanderbijlpark, South Africa
nqume@vut.ac.za
ORCID: 0000-0002-8275-2016

Tranos Zuva
Vaal University of Technology
Computer Sciences and Engineering
Vanderbijlpark, South Africa
tranosz@vut.ac.za
ORCID: 0000-0001-9579-3899

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026

P. Ramuhovhi, N. Sonhera, and T. Zuva
"Enhancing Cybersecurity with Random Forest: Efficient Detection of Cyberattacks",
Latin-American Journal of Computing (LAJC), vol. 13, no. 2, 2026.

Enhancing Cybersecurity with Random Forest: Efficient Detection of Cyberattacks

Phathutshedzo Cyprin Ramuhovhi
Vaal University of Technology
Computer Sciences and Engineering
Vanderbijlpark, South Africa
214102653@edu.vut.ac.za

Naume Sonhera
Vaal University of Technology
Computer Sciences and Engineering
Vanderbijlpark, South Africa
nqume@vut.ac.za

Tranos Zuva
Vaal University of Technology
Computer Sciences and Engineering
Vanderbijlpark, South Africa
tranosz@vut.ac.za

***Abstract*—** The rapid increase in the number of cyberattacks in the digital age has reduced the effectiveness of conventional cybersecurity systems. Traditional methods of cybersecurity face considerable difficulty when detecting new, sophisticated attacks and advanced exploitation techniques swiftly. This research addresses critical cybersecurity concerns by developing an AI-driven Intrusion Detection System (IDS), which employs Random Forest (RF) algorithms to detect cyberattacks efficiently. The evaluation of the model was conducted using three publicly available datasets: CICIDS2017 (692,703 records), NSL-KDD (148,517 records), and UNSW-NB15 (257,673 records) with various attack backgrounds and network configurations. A set of evaluation metrics, including accuracy, precision, recall, and F1-score, was employed to assess the performance of the cyberattack detection prototype. Across the three datasets, the model attained an average accuracy of 99.85%, precision of 99.83%, recall of 99.91%, and an F1-score of 99.87%, while maintaining low error rates, with an average false positive rate of 0.25% and a false negative rate of 0.10%. The results indicate that Random Forest is an effective solution for cyberattack detection in data-driven environments. The model was developed with lightweight and easy-to-deploy criteria, but the evaluation reported in this study was done under benchmark test conditions. This work improves the effectiveness of machine learning-based intrusion detection systems and serves as a stepping stone for future research on operational and real-time deployment of machine learning-based intrusion detection systems.

***Keywords*—** *Artificial Intelligence, Cybersecurity, Cyberattack, Intrusion Detection System*

I. INTRODUCTION

The rapid growth of the digital environment has elevated cybersecurity to a critical priority for both individuals and organizations. The frequency of cyberattacks has increased drastically over the last few decades to keep up with ever-changing technologies [1]. This increase requires organizations to implement strong cybersecurity measures. A study by [1] noted that traditional cybersecurity approaches, such as computer security and network protection systems, are becoming increasingly ineffective in combating continuously evolving and creative cyberattacks. Cyberattacks result in damaged reputations, financial losses due to the theft of intellectual property, legal liabilities, and business operation disruptions [2]. A cyberattack is a deliberate attempt by malicious actors to gain unauthorized access, disrupt operations, or compromise information systems, to destroy data, or perform any malicious activities that will compromise the company network or infrastructure [3]. In the study by [4], traditional cybersecurity is defined as methods such as conventional Intrusion Detection Systems (IDS) that operate based on signature-based and rule-based detection mechanisms in which cyberattacks are detected by matching the data against known attack patterns. The authors further emphasize that these methods are slowly becoming ineffective due to evolving, increasingly sophisticated cyberattacks, often supported or enhanced by Artificial Intelligence (AI) techniques and other advanced cyber capabilities. The world is experiencing an increase in cyberattacks, and the cost is estimated to go up to 10.5 trillion by 2025, compared to 3 trillion in 2015 [5]. In Africa, cases have increased by 76% in 2023, and South Africa has been the most affected by financial losses despite the high-level security measures [6]. Conventional cybersecurity methods, which rely on signature detection, are weak at dealing with advanced threats such as zero-day and polymorphic attacks [1]. Banks remain primary targets for cyberattacks [7], [8], [9] and the fast development of digital systems has increased vulnerabilities. According to recent research, AI-based models, such as SVMs and LSTMs, can increase detection through predicting and simulating attacks [2]. Nonetheless, most AI products are memory-consuming and cannot be used in real-time. The paper fills these gaps by suggesting an optimized, interpretable, and lightweight Random Forest (RF)-based IDS, which is meant to be used in real-time detection in resource-constrained settings, especially in areas such as South Africa, where cyberattacks have dire economic implications.

II. PROBLEM STATEMENT

Most businesses rely on digital technologies, which have resulted in a rise in advanced cyberattacks that conventional cybersecurity approaches fail to detect [2]. The same authors add that conventional cybersecurity approaches involve signature-based and rule-based detection models, which detect cyberattacks by matching data to known attack patterns. These methods often fail to detect new attacks. The complexity of cybersecurity attacks is constantly increasing, presenting serious difficulties for people and companies globally [10]. Traditional IDSs are critical security components that monitor and analyze system activity to detect suspicious behavior and prevent unauthorized access [4]. Research, such as that conducted on [6] and [10], points to AI models as more efficient than traditional algorithms. Deep Learning (DL) and machine learning (ML) models can achieve high accuracy level for real-time detection of zero-day attacks and advanced types of malware. The authors added

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.08

that DL models can achieve high accuracy, but they have high computational requirements. These studies show that traditional cybersecurity methods face significant challenges in detecting sophisticated attacks and advanced exploitation techniques in a timely manner. Thus, this paper delves into using AI to detect cyberattacks in real-time, aiming to improve accuracy and adaptability to emerging cyberattacks.

III. Related Works

Conventional IDSs are based on rule-based and signature-based techniques to identify known attack patterns, which are becoming less effective against contemporary and sophisticated threats [1], [4]. To address these shortcomings, AI technologies are being explored for real-time cyberattack detection, leveraging techniques such as anomaly detection, pattern recognition, and continuous learning[11], [12]. With the ability to provide adaptive and proactive defense, AI can provide the opportunity to identify and react to changing attacks in real-time, a shift from traditional static, signature-based security systems to dynamic, intelligent cybersecurity solutions.

A. The Emergence and Integration of Artificial Intelligence in Various Sectors

AI-based technologies, including machine learning, computer vision, and Convolutional Neural Networks (CNNs), are increasingly applied in agricultural settings to support real-time monitoring of plant conditions, detect diseases at the earliest stages, and manage water in irrigation, optimizing water usage and improving productivity [13]. Advanced deep learning has significantly improved the accuracy of medical image analysis, supporting the diagnosis of conditions such as pneumonia, tumors, and COVID-19, and AI-enabled virtual care systems, based on wearables and chatbots, enable continuous health monitoring and personalized treatment in healthcare [14]. Likewise, in education, AI promotes adaptive learning by having intelligent tutoring systems and automates grading, applying natural language processing (NLP) techniques to support automated grading. These technologies contribute to more personalized learning experiences while allowing educators to focus on innovative teaching practices [15].

AI has been shown to have a lot of transformative capability in diverse fields, such as education, agriculture, and healthcare, where it has increased the efficiency of running operations and decision-making [12], [13], [14]. In the information technology industry, in particular, AI solutions have become more advanced by enhancing complex data analysis rather than simple automation [16]. In this study, there is a special emphasis on the implementation of AI in the context of cybersecurity systems, where its development process is being leveraged to respond to new challenges in technology.

B. The Emergence and Integration of Artificial Intelligence in Cybersecurity

Increasing sophistication and pervasiveness of cyberattacks demand that organizational defense systems change their paradigm [17]. Traditional rule-based security measures are no longer effective when confronted with the adaptive and advanced nature of modern cyberattacks [1]. As a result, improving cybersecurity operations now requires the integration of AI. The main function of AI in enhancing cyber defense, as per the body of existing literature, is critically examined in this section [18].

A study by [2] examined network intrusion detection by testing RF, SVM, LSTMs, and Autoencoders using the CICIDS2017 and UNSW-NB15 datasets. The research examined model performance when applied to real-world attacks to identify a model that can best detect cyberattacks with improved accuracy and efficiency. The RF model displayed maximum accuracy at 92.3% when compared to 89.7% accuracy achieved by SVM within the ML models. LSTMs achieved 94.1% successful attack detection through their sequential approach at the expense of 200 ms computational time per task. Autoencoder achieved higher detection success than K-Means Clustering because its accuracy rate exceeded 87.8%, whereas K-Means scored at 85.4%. These results suggest that although LSTM achieves higher accuracy, it incurs higher computational costs.

A study conducted by [19] was aimed at enhancing security in satellite-terrestrial integrated networks (STIN) using four hybrid IDSs that integrate advanced feature selection with ML and DL methods. The authors optimized the STIN and UNSW-NB15 dataset feature sets using sequential forward selection based on RF to minimize computational cost while optimizing detection performance. RF-based model reached 90.5% on satellite data and 78.52% on terrestrial traffic, and DL variants incorporating RF feature extraction and Gated Recurrent Unit (GRU) reached 87% and 79%, respectively. Architectures based on Long Short-Term Memory (LSTM) and GRU have also been effective in identifying complex attack patterns, including distributed denial-of-service (DDoS) attacks. However, traditional machine learning approaches, particularly when combined with ensemble methods and multilayer perceptron models, continue to perform competitively in detecting such threats. The results demonstrate the effectiveness of hybrid IDS designs in dealing with changing threats in integrated network environments. Research has shown that machine learning methods can help to detect intrusions, especially to counter DDoS attacks.

In a study by [19], it was stated that RF with preprocessing methods like min-max scaling and outlier detection had an impressive accuracy of 99.72% at identifying DDoS attacks. In evaluation studies, several classifiers, including BayesNet, Naive Bayes, J48, Partial Decision Trees (PART), and Random Forest, have been tested on the NSL-KDD dataset alongside dimensionality reduction techniques such as Principal Component Analysis (PCA) and Random Projection (RP) [20]. The findings showed that RP combined with the PART classifier provided the best performance with a 82.0% accuracy and a false positive rate of 16.2, and the F1 score was equal in normal and anomaly classes (82.3% and 81.7%). Additionally, RP was found to be better than PCA in sustaining the classification accuracy and minimizing the computational cost, which highlights its applicability in intrusion detection systems that require efficiency and accuracy in real time.

Data balancing and feature engineering are critical in the field of cybersecurity. Techniques such as SMOTE are commonly used to address class imbalance, particularly in datasets where malicious instances are underrepresented. Flow-related features (packet size, duration, protocol type) are

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.08

always among the most important indicators of malicious activity [21]. RF is still widely used due to its feature ranking feature and ability to process high-dimensional data. In a recent study by [22], ML models (RF, SVM, Gradient Boosting), anomaly detection algorithms (K-Means, DBSCAN, Autoencoder), and DL models (Transformer, RNNs, CNNs) were compared in terms of real-time attack detection. RF was 94.5% accurate, whereas CNNs were 95.8% accurate, and their cost of computation was lower, so they were the most feasible option. The Transformers had the highest accuracy of 96.2% and required 18.3 GB of RAM and 8.5 hours for training, whereas Autoencoders had the highest true positive rate of 94% and the lowest latency of 120 ms, but the highest resource use. Although deep learning models are very accurate, they require intensive calculations and are not interpretable, which prevents their use in real-time. RF on the other hand provides the best tradeoff between accuracy, efficiency, and explainability and good generalization across datasets and low resource consumption- making it one of the top choices in the next-generation intrusion detection systems.

A significant gap in the research on lightweight model integration into real-time resource-constrained environments is evident. The gap addressed in this paper is the creation of an RF-based IDS that is optimized in terms of speed, interpretability, and efficiency. A comparison (Table I) between the existing methods shows that there are trade-offs between current methods: signature-based systems are only able to detect known threats; state-of-the-art AI models like Transformers and LSTMs are highly accurate, but impractical because of resource requirements; and methods like RP with PART classifiers have high false positive rates. This analogy supports the necessity of light, interpretable, and high-performing models like the Random Forest that can balance the detection with the feasibility of the operation.

TABLE I. LIMITATIONS OF EXISTING CYBERSECURITY APPROACHES

Author/s	Method	Detection Accuracy	Key Limitations	Operational Impact
[4]; [1]	Signature/ Rule-based	Focus on known attacks only	Restricted and limited to known attack patterns because it relies on signatures/rules.	- Increased breach Risk - High remediation costs - System downtime
[20]	Random Projection and PART Classifier	82.0% Attack Detection	Misses 18% of the attacks (breach risk), and with 16.2% it has false positives (alert fatigue).	16.2% false positives, SOC teams spend time on a false sense of security and miss real attacks.
[10]	Transforme r	96.2% Attack Detection	Heavy Resource Requirements: 18.3 GB RAM, 8.5h train time.	Cost-prohibitive to the resource-limited organisation.
[2]	LSTM	94.1% Attack Detection	The model's major drawback is its 200 ms computational time, which is a significant cost.	Extra computing costs cause more operational expenses.

The comparative analysis highlights the weaknesses of both traditional and modern approaches to cybersecurity: deep learning models, despite their high accuracy, are costly in terms of resources and cannot be applied in real-time, whereas lightweight models tend to lose their accuracy, leading to alert fatigue. Such gaps highlight the need for a solution that is correct, computationally efficient, and interpretable. To overcome this, the present study proposes an RF-based IDS that is optimized to perform real-time detection in resource-constrained environments.

IV. THE MODEL FRAMEWORK, ARCHITECTURE, AND PROPOSED METHODOLOGY

To improve cyberattack detection with high accuracy, robustness, and interpretability, this paper proposes a Random Forest-based IDS model. The framework was implemented in Python, with a Streamlit interface for model training, test data uploading, attack detection, model persistence, and result logging. The implementation also incorporates machine learning and pre-processing libraries, such as Pandas, NumPy, Scikit-learn, Imbalanced-learn, Matplotlib, and Seaborn.

A. System Architecture

The layers of architecture include the following:

- **Data Acquisition Layer**
  - Gathers unprocessed network traffic content of benchmark data: CICIDS2017, NSL-KDD, UNSW-NB15.
  - The datasets were selected due to varying traffic behaviors and attack types, allowing for the model's performance to be assessed in varying cybersecurity environments.
- **Data Preprocessing Layer**
  - Standardizes data (eliminates duplicates, blank values).
  - Compatibility between datasets was achieved by standardizing the target column across all datasets, identified dynamically using labels like label, attack cat, attack, class, and target.
  - One Hot Encoding and Min-Max Normalization were performed on categorical and numerical variables, respectively. The preprocessing tasks were implemented using: OneHotEncoder, MinMaxScaler, SimpleImputer, ColumnTransformer, and Pipeline.
  - These datasets were then split into training (70%), validation (15%), and test (15%) sets for model development and evaluation.
  - Imbalance was addressed by employing balancing techniques, specifically SMOTE, during the training process when class imbalance was detected, but not in the independent data used for evaluation.
  - Pre-processing and interpretation were handled with caution for features that were not representative of meaningful network behavior, in particular identifier-type

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.08

features, to avoid adding record-specific artefacts to the model that distracted from actual traffic-related features.

- **Feature Engineering Layer**
  - Performs feature selection using Random Forest's in-built importance ranking.
  - Computes Gini impurity to assess feature relevance.
  - The most relevant features were selected based on their importance scores to improve model efficiency, focusing on specific traffic characteristics.
  - The selected features were retained throughout the implemented workflow until approximately 95% cumulative importance had been captured.
- **Model Development Layer**
  - The Random Forest classifier was chosen as the primary detection method given its efficiency, interpretability, and robustness in dealing with high-dimensional data.
  - Uses ensemble voting by multiple decision trees.
  - Hyperparameter optimization was performed by using RandomizedSearchCV to enhance the performance of the model when training.
  - The parameter search space included the number of estimators (n_estimators = 100, 200, 300, 400, 500), maximum depth (max_depth = 15, 20, 30, 40, 50), and minimum samples required for split (min_samples_split = 2, 5, 10). The other parameters used were bootstrap=True, max_features='sqrt', criterion='gini', and random_state=42.
  - The optimization process employed n_iter=10, cv=5, scoring='accuracy' and n_jobs=4, and the best estimator found during the optimisation process was chosen as the resulting model.
- **Model Training Layer**
  - After preprocessing and feature selection, the training subset was used for model fitting.
  - The training workflow was handled in the application environment, and the progress of the models was tracked on a training status file and execution logs.
  - If class balancing was applied, only the selected training features were chosen to undergo SMOTE and random_state=42 was applied. Balanced training subset was then used for Random Forest training and hyperparameter optimization.
  - The validation subset was tracked for model behavior and the test subset was kept separate for final evaluation. This made it possible to evaluate the model on previously unseen records.
- **Evaluation Layer**
  - Evaluates model using Accuracy, Precision, Recall, and F1-Score metrics.
  - The measures were employed to assess the model's overall classification performance and the model's behavior with respect to false positive and false negative items.
  - Validation and test metrics were stored in specific result folders to ensure that the outcome would be traceable and even reusable.
- **Detection Layer**
  - The trained model was then used to detect cyberattacks on uploaded or pre-saved test data via the interface of the system.
  - Application also saved the trained model, preprocessing object, selected feature information, test metrics, validation metrics, and execution logs for re-use.
  - The design of the framework was therefore kept lightweight and deployment-friendly. But the evaluation conducted in this study was primarily based on benchmark datasets rather than real-time deployment.

B. Datasets

In this paper, three benchmark datasets are employed: the CICIDS2017, the NSL-KDD, and the UNSW-NB15. These datasets were chosen as they have a wide variety of both benign and malicious traffic characteristics, allowing the model to be tested in various intrusion detection situations. The validation of the proposed model was also improved by using three known benchmark datasets. The CICIDS2017 dataset was used because it has realistic enterprise network traffic with both benign and attack records. The NSL-KDD dataset was selected as it is still a well-known dataset for intrusion detection studies and provides several categories of attacks. The UNSW-NB15 dataset was added as it has more recent and more complex attack patterns which are relevant for the modern cybersecurity evaluation. Each data set had its target variable selected during the data pre-processing. After which the data sets were cleaned, encoded, scaled, and further reduced through feature selection using Random Forest feature importance. This enabled the selection of the most relevant variables to be used when training and evaluating the model.

C. Model Evaluation Using Benchmark Datasets

Random Forest is a technique of ensemble learning that constructs a variety of decision trees and combines their predictions to enhance generalization and combat overfitting [22]. The RF-based IDS was evaluated using three benchmark datasets: NSL-KDD (148,517 records, 31

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.08

features), CICIDS2017 (692,703 records, 37 features), and UNSW-NB15 (2,540,044 records, 49 features).

- **Mathematical Formulations:**
  - **Preprocessing**

Encoding Stage: The categorical features (such as protocol type and service) went through One-Hot Encoding.

Scanning -Min-max normalization was used to scale the numerical variables to a range of 0 to 1. The normalized value (scaled) of X represents X_{norm} Following the implementation of Min-Max scaling, represented by equation 1.

$$X_{norm} = \frac{X - X_{Min}}{X_{Max} - X_{Min}} \quad (1)$$

where:

X: The original value of a feature

X_{min}: The feature lowest value in the dataset

X_{max}: Highest feature value in the dataset

Feature Importance Formula (Gini-Based): The RF algorithm has calculated the feature importance by comparing the amount of contribution of each feature to better classification accuracy. The features that invariably resulted in larger reductions of impurity were allocated greater importance, indicating that they had a greater impact on the predictive ability. $\text{Importance}(X_m)$ as shown in Equation 2 [23].

$$Importance(X_m) = \frac{1}{B}\sum_{b=1}^{B} \Delta impurity\,(X_m, T_b) \quad (2)$$

where:

B: The total number of trees
Tb: The b-th decision tree in the forest
Δimpurity: Gini decrease from splitting X_m in tree T_b

- **Prediction**

A prediction of ŷ emerges by combining the votes of several decision trees. This ensemble approach enhanced the robustness and accuracy of the final classification, as shown in equation 3 [24]:

$$\hat{y} = mode\left(T^1(x), T^2(x), \ldots, T_n(x)\right) \quad (3)$$

where:

The $T_i(x)$ term represents the likelihood of the i-th decision tree.

- **Split Evaluation - Gini Impurity:**

RF employed Gini Impurity as its measurement tool to determine the split quality of data features. $G(p)$ is the Gini Impurity of the dataset, as presented in equation 4 [24]:

$$G(p) = 1 - \sum_{i=1}^{c} Pi^2 \quad (4)$$

where:

C = class count
p_i = likelihood of class I in the dataset p

D. *Evaluation Metrics*

This sub-section indicates that a rigorous test was conducted to evaluate the model's ability to detect attacks. A list of conventional evaluation metrics was employed to obtain a comprehensive understanding of the competence of the model.

TP = True Positives, TN = True Negatives

FP = False Positives, FN = False Negatives

The following metrics were used for evaluation:

- **Accuracy** -the percentage of cases, both true positives and true negatives, that were correctly classified out of all the instances shown by equation 5 [2].

$$\boldsymbol{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \quad (5)$$

- **Precision and Recall** -Recall shows the number of found anomalies among all detected abnormalities, and precision calculates the proportion of these from all identified items, as shown by equations 6 and 7, respectively [2].

$$\boldsymbol{Precision} = \frac{TP}{TP+FP} \quad (6)$$

$$\boldsymbol{Recall} = \frac{TP}{TP+FN} \quad (7)$$

- **F1-Score** -The one metric that strikes a balance between precision and recall is the harmonic mean of the two, shown by equation 8 [2].

$$\boldsymbol{F1-Score} = 2 \times \frac{Precision \times Recal}{Precision + Recall} \quad (8)$$

E. *Implementation and Reproducibility Details*

The study was done in Python with a Streamlit application. Libraries used for the implementation were pandas, NumPy, joblib, matplotlib, seaborn, scikit-learn, and imbalanced-learn. The key machine learning algorithms used were Random Forest Classifier, train_test_split, RandomizedSearchCV, One Hot Encoder, Min Max Scaler, SimpleImputer, Column Transformer, Pipeline, and SMOTE. Hyperparameter optimization using RandomizedSearchCV was used to train the Random Forest model. The search space consisted of n_estimators (100, 200, 300, 400, 500), max_depth (15, 20, 30, 40, 50), min_samples_split (2, 5, 10), and other parameters like bootstrap=True, max_features='sqrt',criterion='gini', and random_state=42. The best-performing estimator for the optimization process was picked with the n_iter=10, cv=5, scoring='accuracy', and n_jobs=4 option. The application was designed to generate separate folders for data extract processing, model execution, model validation results, test results, and system logs. It also stored the trained model (rf_model.pkl), preprocessing object (preprocessor.pkl), selected feature indices (selected_features.npy), test and validation metrics, backup test data, and training-status records. All of these design decisions facilitated a smooth flow of execution, the reuse of trained artefacts, and the repeatability of the executed workflow. The data preprocessing, feature selection using Random Forest importance, balancing of training data if needed, model training, and evaluation were performed according to the experimental workflow, and the standard evaluation

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.08

criteria were used: accuracy, precision, recall, F1-score, and confusion matrix. This workflow allowed the provided results to be reported through a structured and reproducible machine learning workflow.

V. RESULTS

This section provides an overview of the empirical results obtained from applying the Random Forest-based model to three widely used benchmark cybersecurity datasets: CICIDS2017, NSL-KDD, and UNSW-NB15. Individual datasets are subjected to preprocessing and feature engineering to align with the objective of the study, which is to develop an efficient cyberattack detection model. This section aims to discuss the results of the model in differentiating between benign and malicious traffic on various datasets.

A. CICIDS2017 Dataset Results

The CICIDS2017 dataset is large in scale, containing approximately 11.8 million network flow instances and 85 features [25]. It contains both normal network traffic and various forms of malicious traffic gathered during several days, which resembles the real-life environment of an enterprise [25]. Within the framework of this study, a specific subset of the CICIDS2017 files was selected: WorkingHours.pcap_ISCX. The dataset used contained 692,703 records. It was selected because it is representative and comprises both normal and attack traffic in a real-world environment, as explained by [25]. This subset provides a balanced representation of daily network activity while reducing computational overhead.

1) Model Performance Metrics

The RF-based classifier performed remarkably well on the CICIDS2017 dataset, as can be observed in the confusion matrix shown in Table II. The confusion matrix explains the performance of the model on the test set, and it had 65,988 true negatives and 37,895 true positives. Importantly, it recorded 6 false negatives, which is equivalent to 0.026% false negatives, as well as 17 false positives, which is equivalent to 0.016% false positives. The low number of false positives and false negatives indicates strong classification performance.

TABLE II. CONFUSION MATRIX FOR THE CICIDS2017 DATASET, SHOWING TRUE/FALSE

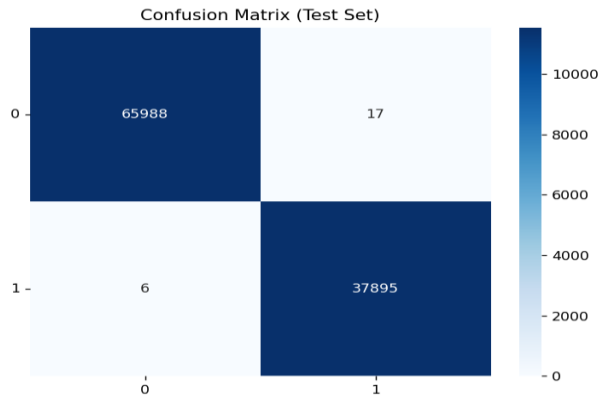


Table III below depicts the model evaluation matrices and results on the CICIDS2017 dataset. The model achieved a classification accuracy of 99.98%, accompanied by a precision of 99.96% and a recall 99.98%. The consistent effectiveness between false positives and false negatives is further supported by an F1 score of 99.97%, showing a balanced performance. The findings reveal that the model was able to successfully classify benign and malicious traffic in the tested CICIDS2017 subset.

TABLE III. RANDOM FOREST MODEL PERFORMANCE RESULTS ON CICIDS2017 DATASET

Metric	Value
Accuracy	99.98%
Precision	99.96%
Recall	99.98%
F1 Score	99.97%

Fig. 1 shows the feature importance rankings for the CICIDS2017 dataset to the RF model. The feature importance results demonstrated that the model was not only based on individual data points but also on the behavior of traffic patterns, with several traffic-related variables influencing classification.

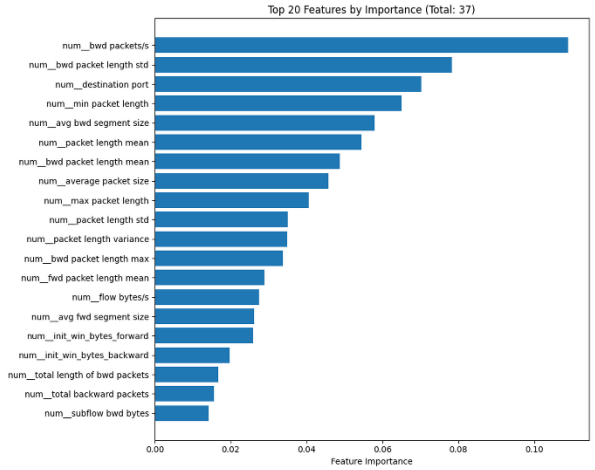


Fig. 1. Feature importance rankings for the CICIDS2017 dataset, derived from the Random Forest model.

B. NSL-KDD Dataset Results

The NSL-KDD dataset is an improved version of the original KDD dataset, KDD-99, which was specially developed to overcome the problems of duplicate and imbalanced nature that plagued the earlier one [26]. It includes four main types of attacks, namely DDoS, Probe, R2L, U2R, and normal network traffic [26]. The NSL-KDD dataset was selected as a benchmark as it contains a standard evaluation dataset for intrusion detection research and multiple categories of attacks with normal traffic. This enabled the model to be tested with well-known intrusion patterns.

1) Model Performance Metrics

For the NSL-KDD dataset, the confusion matrix revealed 11,545 true negatives and 10,708 true positives, indicating that the model successfully classified most of the normal and malicious samples. The model generated 14 false positives and 11 false negatives, indicating strong detection performance.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.08

TABLE IV. CONFUSION MATRIX FOR THE NSL-KDD DATASET, SHOWING TRUE/FALSE POSITIVES/NEGATIVES

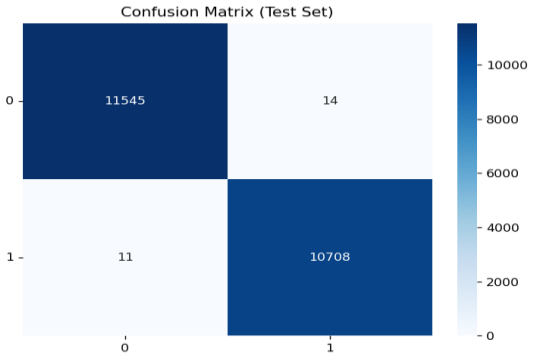


Table V. below depicts the model evaluation matrices and results on the NSL-KDD dataset. The results of the evaluation were a classification accuracy of 99.89%, with a precision of 99.87%, going just a bit higher than the recall of 99.90%. This correlation has produced a balanced F1 score of 99.88%, meaning that both attack detection and false alarm prevention have been performed consistently. The values are preserved throughout the reported results, and it can be seen that they are an unambiguous measure of model performance in the NSL-KDD dataset.

TABLE V. RANDOM FOREST MODEL PERFORMANCE RESULTS ON NSL-KDD DATASET

Metric	Value
Accuracy	99.89%
Precision	99.87%
Recall	99.90%
F1 Score	99.88%

Fig. 2 shows the feature importance for the NSL-KDD dataset.

The feature importance analysis revealed that the traffic-related rate, num_src_bytes, and num_dst_bytes are important for classification. These features are true traffic behavior features and so are applicable in intrusion detection.

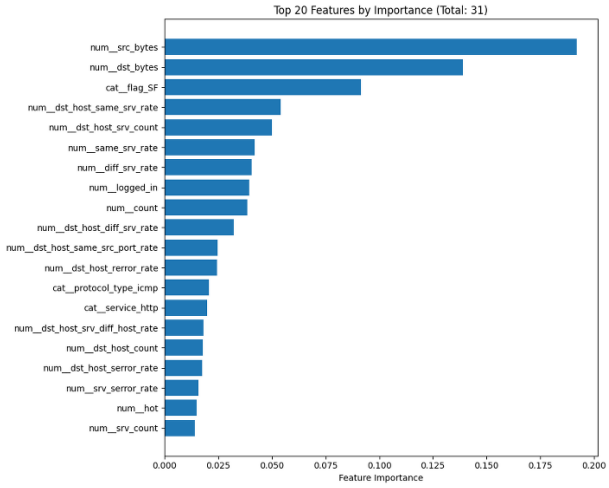


Fig. 2. *Feature importance rankings for the NSL-KDD dataset, derived from the Random Forest model*

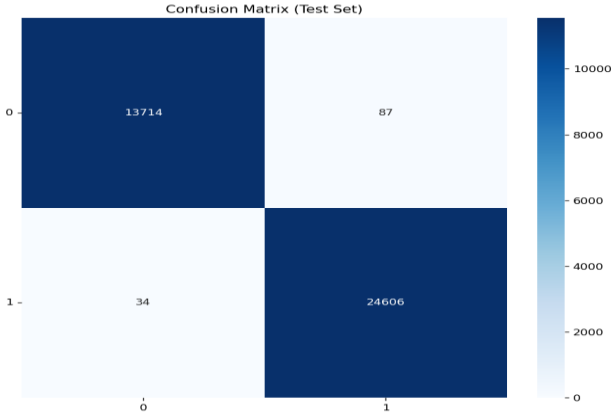
C. UNSW NB15 Dataset Results

The UNSW-NB15 dataset offers recent network traffic, including state-of-the-art types of attacks, meaning Exploits, Fuzzers, and Shellcode [27]. The UNSW-NB15 was generated with IXIA Perfect Storm, a professional network traffic generator, which exhibits high variability, thus making it a strenuous and realistic benchmark to test IDSs [28]. The entire dataset of 257,673 records was utilized for analysis, and 33 features were used.

1) *Model Performance Matrices*

The confusion matrix of the model on the test set, depicted in Table VI, gives a quantitative measure of the performance of the model concerning its classification. The confusion matrix revealed 13,714 true negatives and 24,606 true positives, highlighting good classification accuracy for both benign and malicious traffic. The model achieved a low number of 87 false positives and a low number of 34 false negatives, among the correctly classified instances.

TABLE VI. CONFUSION MATRIX FOR THE UNSW NB15, SHOWING TRUE/FALSE POSITIVES/NEGATIVES



The RF model provided good detection performance in the modern UNSW-NB15 dataset, as depicted in Table VII. below. The overall accuracy of the classifier was 99.69%, the precision was 99.65%, the recall was 99.86%, and the F1 score of 99.75%. The values are slightly lower than those found in CICIDS2017 and NSL-KDD, as attack patterns are far more complex and varied in the UNSW-NB15 dataset.

TABLE VII. RANDOM FOREST MODEL PERFORMANCE RESULTS ON UNSW NB15 DATASET

Metric	Value
Accuracy	99.69%
Precision	99.65%
Recall	99.86%
F1 Score	99.75%

Fig. 3 shows the feature importance rankings for the UNSW-NB15 dataset. The feature importance analysis showed that the traffic related attributes namely sttl and ct_state_ttl were significant features affecting the classification performance. Identifier-based variables, like num_id, were however interpreted with a certain caution as they do not describe intrinsic network behaviour and might be misinterpreted as an indexing effect and not as a significant attack characteristic.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.08

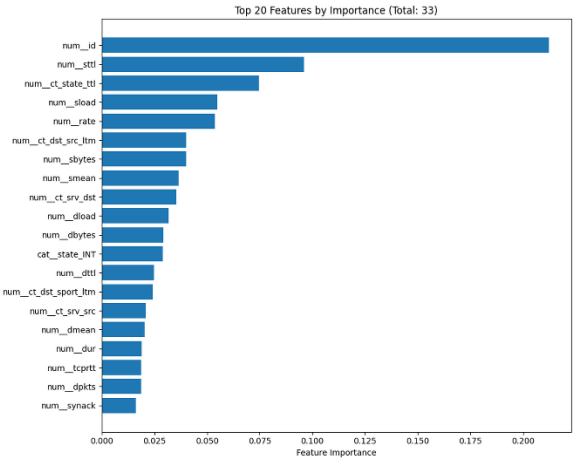


Fig. 3. *Feature importance rankings for the UNSWNB15 dataset, derived from the Random Forest model*

D. Summary of Results Across Datasets

In all three datasets, the model based on the Random Forest algorithm showed consistently high classification performance. CICIDS2017, NSL-KDD, and UNSW-NB15 were found to be the top three performing datasets. The slightly lower performance in the UNSW-NB15 data set is due to the fact that the patterns are more complex and newer. In all the datasets, the confusion matrices exhibited low false positive and false negative rates, demonstrating that the model could maintain a balance between detecting malicious activity and giving few false alarms in the benchmark set used. Overall, the results show that the Random Forest model is effective for cyberattack detection within the experimental setup used in this study. The results discussed here are in line with its suitability to be a lightweight and practical detection approach, but do not fully represent a real-time, fully operational deployment.

VI. DISCUSSION

The findings of this study demonstrate that the Random Forest model can be effectively used for cyberattack detection in a variety of datasets. Its high generalizability and reliability are reflected in accuracy, precision, and recall values exceeding 99%. The minor difference in the performance of the datasets, especially the reduced accuracy of the UNSW-NB15 dataset when compared to CICIDS2017 and NSL-KDD, is due to the complexity and stealthiness of the attacks in the UNSW-NB15 dataset. This is consistent with literature results that current data sets with sophisticated types of attacks are more difficult to detect using detection systems.

On the CICIDS2017 dataset, the current model achieved an accuracy of 99.98%, precision of 99.96%, recall of 99.98%, and an F1-score of 99.97%, based on 692,703 records and 37 features. Comparatively, a study by [27] achieved an equally good performance of 99.94% accuracy, 99.94% precision and recall, and 99.94% F1-score with only 10 features. A study by [2] shows inferior results on every measure (92.3% accuracy, 90.5% precision, 88.2% recall, and 89.3% F1-score), and the study employed 80 features.

The same trend can be observed in the analysis of the UNSW-NB15 dataset. A subset of 257,673 records and 33 features was used in the current paper, and the detection accuracy is 99.86%, 99.65% precision, 99.86% recall, and a 99.75% F1-score. A study by [29] was conducted based on the full dataset containing 2,540,044 records, whereas the present one was based on a sample of the subset of 257,673 records. The dataset size was not indicated by [30], while the study conducted by [31] used a subset of 257,673, which aligns with the dataset used in the current study. The number of selected features ranged from 19 in [31] to 49 in [29], while the current study used 33 features. As the performance matrices indicate, study by [31] achieved an accuracy of 95.05%, which is the lowest in the comparison. The results reported in [29] were significantly high and comprised 99.42% accuracy, 99.71% precision, 99.63% recall, and 99.67% F1-score. The current model achieved comparable performance to the studies reviewed with 99.86 % accuracy, 99.65 % precision, 99.86 % recall, and 99.75 % F1-score. The study by [30] achieved 98.7% accuracy, but did not provide other performance indicators, which is why the comparison cannot be considered complete.

On the NSL-KDD dataset, the current model showed once again superior performance with 99.89% accuracy, 99.87% precision, 99.90% recall, and 99.88% F1-score on the full dataset of 148,517 records and 31 features. The matrices of evaluation indicate that [19] achieved a high accuracy of 99.72%, an F1-score of 99.70%, the highest precision of 99.84%, and a recall of 99.56%. The current study has been found to compete effectively with the highest accuracy of 99.89%, precision of 99.87%, recall of 99.90%, and an F1-score of 99.88%. In a study by [32] had a bit lower performance in all matrices (98.75-98.76%), although the number of features used by them was the lowest. Study by [30] did not present other performance measures for comparison with the high accuracy of 99.65%. Overall, the comparative analysis highlights the strength of the existing model in a wide range of datasets and experimental states. Simultaneously, it shows significant discrepancies in the manner the studies report dataset size, feature selection, and evaluation metrics. Such discrepancies make comparisons difficult and underline the necessity of developing the standardized evaluation frameworks within the research of intrusion detection. These frameworks would help in reproducibility, transparency, and more credible benchmarking of future models.

VII. CONCLUSION

This work presented and tested a Random Forest-based Intrusion Detection System (IDS) for the detection of cyberattacks using three benchmark datasets, namely CICIDS2017, NSL-KDD, and UNSW-NB15. The model was created to be lightweight and efficient for distinguishing network traffic as benign or malicious. The results demonstrate that the Random Forest model achieves high classification accuracy across different datasets in various traffic conditions. The evaluation results showed that the model performance was balanced in terms of accuracy, precision, recall, and F1-score, and that the number of false positives as well as the number of false negatives in the evaluated data sets were kept at a low level. These results show how well the model can differentiate between normal and attack traffic in various circumstances. It was also noted that the model was consistent across the various datasets, with slightly poorer performance on the UNSW-NB15 dataset, as

the attacks were more complex and varied. This indicates that dataset complexity plays an important role in the intrusion detection performance of the model. The feature analysis revealed that traffic-related features provided more relevant information for classification than identifier-based features. These findings highlight the importance of using network-related features that are relevant to actual network behavior for the development of intrusion detection systems. The study shows the effectiveness and interpretability of the Random Forest algorithm for cyberattack detection within the scope of the experimental evaluation carried out. Furthermore, the application of a structured preprocessing and evaluation pipeline and multi-dataset validation results in robust findings. The model was designed for deployment-oriented attributes, while the evaluation in this study was conducted in a benchmark-based environment. The results, therefore, do not necessarily reflect the capability for detection in a fully validated real-time operation, but rather under controlled experimental conditions. This study does not directly compare with deep learning models like LSTM and RNN under the same experimental conditions. For this reason, it is not claimed that the above models are superior. Rather, the results indicate that Random Forest-based approach is a practical solution in terms of performance, interpretability, and computational requirements for intrusion detection problems. Future research should be conducted to test the model in a fully operational setting, measuring inference latency, throughput, memory consumption and processing capacity in real-time. Moreover, future studies can investigate the use of hybrid models using a combination of Random Forest and additional machine learning or deep learning models for more accurate detection of sophisticated and evolving cyberattacks.

REFERENCES

[1] F. Tao, M. Akhtar, and Z. Jiayuan, "The future of Artificial Intelligence in Cybersecurity: A Comprehensive Survey," EAI Endorsed Trans. Creat. Technol., vol. 8, no. 28, p. 170285, Aug. 2021, doi: 10.4108/eai.7-7-2021.170285.

[2] V. Jain and A. Mitra, "Real-Time Threat Detection in Cybersecurity: Leveraging Machine Learning Algorithms for Enhanced Anomaly Detection," in Advances in Computational Intelligence and Robotics, M. A. Almaiah and Y. Maleh, Eds., IGI Global, 2024, pp. 315–344. doi: 10.4018/979-8-3693-7540-2.ch014.

[3] Ö. Aslan, S. S. Aktuğ, M. Ozkan-Okay, A. A. Yilmaz, and E. Akin, "A Comprehensive Review of Cyber Security Vulnerabilities, Threats, Attacks, and Solutions," Electronics, vol. 12, no. 6, p. 1333, Mar. 2023, doi: 10.3390/electronics12061333.

[4] M. Markevych and M. Dawson, "A Review of Enhancing Intrusion Detection Systems for Cybersecurity Using Artificial Intelligence (AI)," Int. Conf. Knowl.-BASED Organ., vol. 29, no. 3, pp. 30–37, June 2023, doi: 10.2478/kbo-2023-0072.

[5] Seacom, "Financial impact of security breaches is highest in SA," Jan. 04, 2024. [Online]. Available: https://seacom.co.za/news/financial-impact-of-security-breaches-is-highest-in-sa

[6] Aswa, "AI-Powered Cybersecurity: Leveraging Deep Learning for RealTime Threat Detection and Prevention," 2025, 2025, doi: 10.18535/ijecs/v14i01.4975.

[7] Ayodeji Oyindamola Ikudabo, Chinedu C. Onyeje, Daniel O. T. Ihenacho, and K. C. Nwafor, "Mitigating cybersecurity risks in financial institutions: The role of AI and data analytics," Int. J. Sci. Res. Arch., vol. 13, no. 1, pp. 2895–2910, Oct. 2024, doi: 10.30574/ijsra.2024.13.1.2014.

[8] A. Khemka, "The impact of cyber attacks on financial institutions and the need for improved security measures.," vol. 9, no. 10, 2024.

[9] Md Anwarul Matin Jony, Rashedul Islam, and F. H. Muhammad Saqib Jalil, "AI-Powered Cybersecurity in Financial Institutions: Enhancing Resilience Against Emerging Digital Threats," Adv. Int. J. Multidiscip. Res., vol. 2, no. 6, p. 1113, Nov. 2024, doi: 10.62127/aijmr.2024.v02i06.1113.

[10] H. Hussain, M. Kainat, M. Tunio, and T. Ali, "Leveraging AI and Machine Learning to Detect and Prevent Cyber Security Threats," Jan. 2025, doi: 10.5281/ZENODO.14714679.

[11] N. Katiyar, S. Tripathi, Mr. P. Kumar, Mr. S. Verma, A. K. Sahu, and S. Saxena, "AI and Cyber-Security: Enhancing threat detection and response with machine learning.," Educ. Adm. Theory Pract., Apr. 2024, doi: 10.53555/kuey.v30i4.2377.

[12] M. Mohammed, A. J. Mohammed, U. U. M. Mohammed, and Z. A. Mohammed, "Advancements in AI-Based Security and Threat Detection," IJARCCE, vol. 13, no. 4, Mar. 2024, doi: 10.17148/IJARCCE.2024.13459.

[13] R. C. D. Oliveira and R. D. D. S. E. Silva, "Artificial Intelligence in Agriculture: Benefits, Challenges, and Trends," Appl. Sci., vol. 13, no. 13, p. 7405, June 2023, doi: 10.3390/app13137405.

[14] A. AI-Kuwaiti et al., "A Review of the Role of Artificial Intelligence in Healthcare," J. Pers. Med., vol. 13, no. 6, p. 951, June 2023, doi: 10.3390/jpm13060951.

[15] C. Meirinhos, L. Fernandes, and M. Meirinhos, "The emergence of Artificial Intelligence in Education," 2023.

[16] K. Meduri, H. Gonaygunt, and G. S. Nadella, "Evaluating the Effectiveness of AI-Driven Frameworks in Predicting and Preventing Cyber Attacks," Int. J. Res. Publ. Rev., vol. 5, no. 3, pp. 6591–6595, Mar. 2024, doi: 10.55248/gengpi.5.0324.0875.

[17] R. Kaur, D. Gabrijelčič, and T. Klobučar, "Artificial intelligence for cybersecurity: Literature review and future research directions," Inf. Fusion, vol. 97, p. 101804, Sept. 2023, doi: 10.1016/j.inffus.2023.101804.

[18] A. O. Adewusi, U. I. Okoli, T. Olorunsogo, E. Adaga, D. O. Daraojimba, and O. C. Obi, "Artificial intelligence in cybersecurity: Protecting national infrastructure - A USA review," World Journal of Advanced Research and Reviews, vol. 21, no. 1, pp. 2263–2275, 2024, doi: 10.30574/wjarr.2024.21.1.0313

[19] I. Avci and M. Koca, "Cybersecurity Attack Detection Model, Using Machine Learning Techniques," Acta Polytech. Hung., vol. 20, no. 7, pp. 29–44, 2023, doi: 10.12700/APH.20.7.2023.7.2.

[20] F. Nabi and X. Zhou, "Enhancing intrusion detection systems through dimensionality reduction: A comparative study of machine learning techniques for cyber security," Cyber Secur. Appl., vol. 2, p. 100033, 2024, doi: 10.1016/j.csa.2023.100033.

[21] M. Luay, S. Layeghy, S. Hosseininoorbin, M. Sarhan, N. Moustafa, and M. Portmann, "Temporal Analysis of NetFlow Datasets for Network Intrusion Detection Systems," Mar. 09, 2025, arXiv: arXiv:2503.04404. doi: 10.48550/arXiv.2503.04404.

[22] H. A. Salman, A. Kalakech, and A. Steiti, "Random Forest Algorithm Overview," Babylon. J. Mach. Learn., vol. 2024, pp. 69–79, June 2024, doi: 10.58496/BJML/2024/007.

[23] D. Koutsandreas and I. Keppo, "Harnessing machine learning algorithms to unveil energy efficiency investment archetypes," Energy Rep., vol. 12, pp. 3180–3195, Dec. 2024, doi: 10.1016/j.egyr.2024.09.009.

[24] Putta Srivani, "Integrating Natural Language Processing with AdaBoost, Random Forest, and Logistic Regression for an Advanced Ensemble-Based Network Intrusion Detection Model," J. Inf. Syst. Eng. Manag., vol. 10, no. 3s, pp. 264–283, Jan. 2025, doi: 10.52783/jisem.v10i3s.386.

[25] Zafar Iqbal Khan, Mohammad Mazhar Afzal, and Khurram Naim Shamsi, "A Comprehensive Study on CIC-IDS2017 Dataset for Intrusion Detection Systems," Int. Res. J. Adv. Eng. Hub IRJAEH, vol. 2, no. 02, pp. 254–260, Feb. 2024, doi: 10.47392/IRJAEH.2024.0041.

[26] Y. Sahli, "comparison of the NSL-KDD dataset and its predecessor the KDD Cup '99 dataset," Int. J. Sci. Res. Manag., vol. 10, no. 04, pp. 832–839, Apr. 2022, doi: 10.18535/ijsrm/v10i4.ec05.

[27] Md. A. Talukder et al., "Machine learning-based network intrusion detection for big and imbalanced data using oversampling, stacking feature embedding and feature extraction," J. Big Data, vol. 11, no. 1, p. 33, Feb. 2024, doi: 10.1186/s40537-024-00886-w.

[28] S. M. Kasongo, "An Advanced Intrusion Detection System for IIoT Based on GA and Tree Based Algorithms," IEEE Access, vol. 9, pp. 113199–113212, 2021, doi: 10.1109/ACCESS.2021.3104113.

[29] S. More, M. Idrissi, H. Mahmoud, and A. T. Asyhari, "Enhanced Intrusion Detection Systems Performance with UNSW-NB15 Data Analysis," Algorithms, vol. 17, no. 2, p. 64, Feb. 2024, doi: 10.3390/a17020064.

[30] J. Note and M. Ali, "Comparative Analysis of Intrusion Detection System Using Machine Learning and Deep Learning Algorithms," Ann. Emerg. Technol. Comput., vol. 6, no. 3, pp. 19–36, July 2022, doi: 10.33166/AETiC.2022.03.003.

[31] I. H. Putro, "Evaluating the Performance of Machine Learning Classifiers for Network Intrusion Detection: A Comparative Study Using the UNSW-NB15 Dataset," Teknika, vol. 14, no. 2, pp. 330–338, July 2025, doi: 10.34148/teknika.v14i2.1276.

[32] A. M. A. Abdullah and K. A. Abood, "Comparative Analysis of Machine Learning Techniques for Intrusion Detection in IoT Networks," Univ. Aden J. Nat. Appl. Sci., vol. 28, no. 2, pp. 53–60, Apr. 2025, doi: 10.47372/uajnas.2024.n2.a05.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: https://doi.org/10.33333/lajc.vol13n2.08

AUTHORS

Phathutshedzo Cyprin Ramuhovhi

Phathutshedzo Ramuhovhi is an IT professional and researcher with a strong background in information technology and applied artificial intelligence. His academic training has focused on cybersecurity, machine learning, and data-driven systems, with a particular interest in the detection and prevention of cyberattacks. He completed his Master's research in intrusion detection systems, where his work explored the use of machine learning models for cyberattack detection across multiple benchmark datasets.

He is currently advancing his research in artificial intelligence and cybersecurity, with a focus on federated learning, explainable artificial intelligence, and intelligent intrusion detection systems. His work integrates machine learning techniques with practical cybersecurity applications to improve detection accuracy and system efficiency.

In addition to his academic work, he serves as a Maintenance & Support Delivery Manager, contributing to IT operations and system support. His expertise in Python programming, data analysis, and machine learning model development reflects a strong integration of theoretical knowledge and practical industry experience.

Naume Sonhera

Dr Naume Sonhera is a Senior Lecturer and Head of the Department of Computer Science within the Faculty of Applied and Computer Sciences at the Vaal University of Technology (VUT), South Africa. She holds a Doctor of Philosophy (PhD) in Information Systems, a Master of Science (MSc) in Computer Science, and a Bachelor of Science in Education (Licentiate Degree) in Mathematics.

She has held various academic and leadership roles throughout her career, including ICT Coordinator, Head of Academics, and Information Technology Manager, and has also served as Acting Campus Principal at one of the university's satellite campuses.

Dr Sonhera is an established researcher with a strong record of impactful research outputs in Computer Science and ICT. Her research interests include Information and Communication Technologies for Development (ICT4D), cloud computing, cybersecurity, cyber threats, cyberbullying, and artificial intelligence.

She is professionally affiliated with the Institute of Information Technology Professionals South Africa (IITPSA) and the South African Institute of Computer Scientists and Information Technologists (SAICSIT), and has received recognition for her contributions to teaching excellence.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026

AUTHORS

Tranos Zuva

Professor Tranos Zuva is a Professor of Computer Science at the Vaal University of Technology (VUT) and serves as the MICT SETA Fourth Industrial Revolution (4IR) Research Chair. He has over 30 years of experience in teaching, research, innovation, and academic leadership, and is widely recognized for his contributions in Artificial Intelligence, Cybersecurity, Data Science, Software Engineering, Digital Transformation, and Emerging Technologies.

He has published more than 200 peer-reviewed journal articles and conference papers and has successfully supervised numerous Master's and Doctoral students. His work has received significant recognition for its impact on both academia and industry.

Professor Zuva actively promotes industry–academic collaboration, innovation, and capacity building. He plays a leading role in advancing 4IR initiatives, digital skills development, and the application of technology-driven solutions to address societal and industrial challenges in South Africa and beyond.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026

LATIN-AMERICAN JOURNAL OF COMPUTING

Published by

Escuela Politécnica Nacional
Facultad de Ingeniería de Sistemas

Quito-Ecuador

https://lajc.epn.edu.ec/
lajc@epn.edu.ec

July 2026

Vol XIII, Issue 2, July 2026

LATIN-AMERICAN
JOURNAL OF
COMPUTING