SASYR Symposium of Applied Science for Young Researchers

# 5th Symposium of Applied Science for Young Researchers

## PROCEEDINGS 2025

July 2 , 2025

# 5th Symposium of Applied Science for Young Researchers

## Proceedings

## SASYR 2025

2 July 2025

CeDRI
Research Centre in Digitalization and Intelligent Robotics

2Ai APPLIED ARTIFICIAL INTELLIGENCE LABORATORY

gecad
Research Group on Intelligent Engineering and Computing for Advanced Innovation and Development

ipvc adit-lab
Applied Digital Transformation Laboratory
Instituto Politécnico de Viana do Castelo

ipb INSTITUTO POLITÉCNICO DE BRAGANÇA

IPCA POLITÉCNICO DO CÁVADO E DO AVE

P.PORTO

ipvc Instituto Politécnico de Viana do Castelo

# Editors

**Florbela P. Fernandes**
Research Centre in Digitalization and Intelligent Robotics (CeDRI)
Instituto Politécnico de Bragança

**Helena Torres**
Applied Artificial Intelligence Laboratory (2Ai)
Instituto Politécnico do Cávado e do Ave

**Pedro Pinto**
Research Group on Intelligent Engineering and Computing for Advanced Innovation and Development (GECAD)
Instituto Politécnico do Porto

**Silvestre Malta**
Applied Digital Transformation Laboratory (ADiT-LAB)
Instituto Politécnico de Viana do Castelo

Instituto Politécnico de Bragança – 2025
Campus de Santa Apolónia
5300-253 Bragança – Portugal
ISBN: 978-972-745-360-3

Book cover: Natália Santos, Instituto Politécnico do Cávado e do Ave

# Welcome

This document presents the proceedings of the 5th Symposium of Applied Science for Young Researchers – SASYR 2025. This scientific event welcomed scientific works on the topics covered by the following four research centers:

– CeDRI (from IPB, Instituto Politécnico de Bragança)
– 2Ai (from IPCA, Instituto Politécnico do Cávado e do Ave)
– GECAD (from IPP, Instituto Politécnico do Porto)
– ADiT-lab (from IPVC, Instituto Politécnico de Viana do Castelo)

The primary objective of SASYR 2025 is to create a welcoming and relaxed environment for young researchers to present their work, discuss recent findings, and explore new ideas. In this way, this event offers an opportunity for the CeDRI, 2Ai, GECAD, and ADiT-lab research communities to leverage synergies and promote collaborations, thereby enhancing the quality of their research.

The SASYR 2025 took place at Instituto Politécnico de Viana do Castelo, Viana do Castelo, Portugal, on 2 July 2025.

The SASYR 2025 Organizing Committee,

Florbela P. Fernandes, Helena Torres, Pedro Pinto, and Silvestre Malta

# Committees

## Organizing Committee

Florbela P. Fernandes, Instituto Politécnico de Bragança
Helena Torres, Instituto Politécnico do Cávado e do Ave
Pedro Pinto, Instituto Politécnico do Porto
Silvestre Malta, Instituto Politécnico de Viana do Castelo

## Advisory Committee

Ana Pereira, Instituto Politécnico de Bragança
António Miguel Cruz, Instituto Politécnico de Viana do Castelo
Carlos Ramos, Instituto Politécnico do Porto
Goreti Marreiros, Instituto Politécnico do Porto
João L. Vilaça, Instituto Politécnico do Cávado e do Ave
Jorge Garcia, Instituto Politécnico de Viana do Castelo
José Lima, Instituto Politécnico de Bragança
Paulo Leitão, Instituto Politécnico de Bragança

## Scientific Committee

Ana Almeida, Instituto Politécnico do Porto
Ana Faria, Instituto Politécnico do Porto
Ana I. Pereira, Instituto Politécnico de Bragança
Ana Madureira, Instituto Politécnico do Porto
Ana Teixeira, Instituto Politécnico do Porto
Ângela Ferreira, Instituto Politécnico de Bragança
António Miguel Rosado da Cruz, Instituto Politécnico de Viana do Castelo
Beatriz Flamia Azevedo, Instituto Politécnico de Bragança
Bertil P. Marques, Instituto Politécnico do Porto
Carla A. S. Geraldes, Instituto Politécnico de Bragança
Carlos Balsa, Instituto Politécnico de Bragança
Carlos R. Cunha, Instituto Politécnico de Bragança
Cátia Alves, Instituto Politécnico do Cávado e do Ave
Clara Bento Vaz, Instituto Politécnico de Bragança
Constantino Martins, Instituto Politécnico do Porto
Daniel Miranda, Instituto Politécnico do Cávado e do Ave
Diogo Martinho, Instituto Politécnico do Porto
Duarte Duque, Instituto Politécnico do Cávado e do Ave
Estela Vilhena, Instituto Politécnico do Cávado e do Ave
Eva Maia, Instituto Politécnico do Porto
Eva Oliveira, Instituto Politécnico do Cávado e do Ave

Fernando Lezama, Instituto Politécnico do Porto
Fernando Monteiro, Instituto Politécnico de Bragança
Fernando Veloso, Instituto Politécnico do Cávado e do Ave
Flávia Pires, Instituto Politécnico de Bragança
Florbela P. Fernandes, Instituto Politécnico de Bragança
Francisco Pereira, Instituto Politécnico do Porto
Gil Lopes, Instituto Politécnico do Porto
Helena Sofia Rodrigues, Instituto Politécnico de Viana do Castelo
Helena Torres, Instituto Politécnico do Cávado e do Ave
Inês Sena, Instituto Politécnico de Bragança
Isabel S. Jesus, Instituto Politécnico do Porto
Ivone Amorim, Instituto Politécnico do Porto
João Borges, Instituto Politécnico do Cávado e do Ave
João Braun, Instituto Politécnico de Bragança
João Carlos Silva, Instituto Politécnico do Cávado e do Ave
João Coelho, Instituto Politécnico de Bragança
João Mendes, Instituto Politécnico de Bragança
João Paulo Teixeira, Instituto Politécnico de Bragança
Joaquim Gonçalves, Instituto Politécnico do Cávado e do Ave
Joaquim Santos, Instituto Politécnico do Porto
Jorge Esparteiro Garcia, Instituto Politécnico de Viana do Castelo
Jorge Ribeiro, Instituto Politécnico de Viana do Castelo
José Henrique Brito, Instituto Politécnico do Cávado e do Ave
José Lima, Instituto Politécnico de Bragança
Luís Conceição, Instituto Politécnico do Porto
Luís Ferreira, Instituto Politécnico do Cávado e do Ave
Luis Gomes, Instituto Politécnico do Porto
Luiz Faria, Instituto Politécnico do Porto
Maria F Pacheco, Instituto Politécnico de Bragança
Maria João Varanda Pereira, Instituto Politécnico de Bragança
Marílio Cardoso, Instituto Politécnico do Porto
Martinha Pereira, Instituto Politécnico do Cávado e do Ave
Murillo Ferreira Dos Santos, Instituto Politécnico de Bragança
Natália Maria De Bessa Rego, Instituto Politécnico do Cávado e do Ave
Nuno Pereira, Instituto Politécnico do Porto
Orlando Sousa, Instituto Politécnico do Porto
Óscar R. Ribeiro, Instituto Politécnico do Cávado e do Ave
Patricia Leite, Instituto Politécnico do Cávado e do Ave
Paula Alexandra Rego, Instituto Politécnico de Viana do Castelo
Paulo Alves, Instituto Politécnico de Bragança
Paulo Costa, Instituto Politécnico de Viana do Castelo
Paulo Matos, Instituto Politécnico de Bragança
Paulo S. Matos, Instituto Politécnico do Porto
Pedro Coutinho, Instituto Politécnico de Viana do Castelo

Pedro Faria, Instituto Politécnico de Viana do Castelo
Pedro Filipe Oliveira, Instituto Politécnico de Bragança
Pedro Morais, Instituto Politécnico do Cávado e do Ave
Pedro Pinto, Instituto Politécnico do Porto
Pedro Rodrigues, Instituto Politécnico de Bragança
Ramiro Barbosa, Instituto Politécnico do Porto
Ricardo Almeida, Instituto Politécnico do Porto
Ricardo Severino, Instituto Politécnico do Porto
Rogério Tavares, Instituto Politécnico de Bragança
Rui Pedro Lopes, Instituto Politécnico de Bragança
Sara Baltazar, Instituto Politécnico de Viana do Castelo
Sara Cruz, Instituto Politécnico do Cávado e do Ave
Sara Paiva, Instituto Politécnico de Viana do Castelo
Silvestre Malta, Instituto Politécnico de Viana do Castelo
Teresa Abreu, Instituto Politécnico do Cávado e do Ave
Tiago Franco, Instituto Politécnico de Bragança
Tiago Pedrosa, Instituto Politécnico de Bragança
Vítor Carvalho, Instituto Politécnico do Cávado e do Ave

**Local Organizing Committee**

Helena Sofia Rodrigues, Instituto Politécnico de Viana do Castelo
Jorge Garcia, Instituto Politécnico de Viana do Castelo
Sara Baltazar, Instituto Politécnico de Viana do Castelo
Silvestre Malta, Instituto Politécnico de Viana do Castelo

**Secretariat / Technical Support**

Beatriz Cabrera, Instituto Politécnico de Bragança
Carla Fontes, Instituto Politécnico de Bragança
Clarisse Pais, Instituto Politécnico de Bragança
Evandro Alves, Instituto Politécnico de Bragança
Francisco Dias, Instituto Politécnico de Viana do Castelo
Gabriel Brandão, Instituto Politécnico de Viana do Castelo
Ivo Ramoa, Instituto Politécnico de Viana do Castelo
Luís Filipe Dias, Instituto Politécnico de Viana do Castelo
Luís Pinto, Instituto Politécnico de Viana do Castelo
Mario Gomes, Instituto Politécnico de Viana do Castelo
Natália Santos, Instituto Politécnico do Cávado e do Ave
Rodrigo Machado, Instituto Politécnico de Viana do Castelo
Rui Filipe Silva, Instituto Politécnico de Viana do Castelo
Tiago Baptista, Instituto Politécnico de Viana do Castelo

# Table of Contents

# Exploiting Trust: How Artificial Intelligence and Deepfakes Shape Social Engineering Attacks

Sérgio Serra[1], Kely Gonzaga[1], and Silvestre Malta[1,2]

[1] Instituto Politécnico de Viana do Castelo, IPVC, Portugal
{sergioserra, kelygonzaga}@ipvc.pt
[2] ADiT-Lab, Instituto Politécnico de Viana do Castelo, 4900-347 Viana do Castelo, Portugal
smalta@estg.ipvc.pt

**Abstract.** The advancement of Artificial Intelligence (AI) has significantly transformed the field of cybersecurity, particularly in the domain of Social Engineering (SE). This paper presents a comprehensive analysis of how generative AI, Large Language Models (LLMs), and deepfake technologies are reshaping SE attacks, allowing unusual levels of personalization, automation, and deception. Through a systematic review of the literature, the study categorizes the main AI-powered attack strategies and examines the psychological, technological, and ethical implications of their use. The research identifies key vulnerabilities exploited by malicious actors and explores current and proposed mitigation strategies, including defenses driven by AI and regulatory initiatives.

**Keywords:** AI-powered social engineering · Generative AI · Large language models LLMs · Deepfake technology · Cybersecurity.

## 1 Introduction

The rapid escalation of AI in recent years has profoundly transformed the digital environment, exponentially increasing the risks associated with cybersecurity, especially in the area of SE. As technological innovation advances, so do the tools employed by cybercriminals, who exploit AI — broadly defined as a machine or system's ability to mimic human behavior [8] — to manipulate, infiltrate, control and compromise personal and institutional security [12].

Traditionally grounded in psychological manipulation, exploiting trust, urgency and fear, SE, which was easier to detect by human eyes, has evolved with the integration of generative AI, LLM, and deepfakes.These technologies automate and scale attacks, allowing highly personalized human-like communications that bypass conventional filters [10] [9]. The widespread use of social media amplifies the surface of the attack, facilitating the extraction of personal data for customized deception [8]. Modern generative tools AI can create adaptive narratives in real time, while deepfake technologies can simulate superrealistic individuals, improving frauds based on impersonation and phishing campaigns [10]. In addition, "in conjunction with techniques such as jailbreaking, prompt injection, and reverse psychology, these models pose a notable threat to the cybersecurity landscape" [10]. This potent combination challenges traditional detection systems and raises urgent concerns regarding privacy, ethics, and proactive security posture.

Despite growing interest in this topic, significant research gaps remain, particularly in relation to the evolving nature of AI powered SE, its societal implications, and the

need for successful mitigation strategies [12]. Addressing these gaps, this study reviews the current literature, maps technological advances, and examines forward-looking frameworks for detection and defense. The paper is structured as follows: Section 2 details the methodology; Section 3 explores the technological foundations of AI in SE; Section 4 examines key empirical and technical models; Section 5 addresses ethical and regulatory issues; and Section 6 summarizes the findings and outlines future directions.

## 2 Methodology

To understand how AI has been used to improve SE attacks, a systematic review of the literature was carried out. The review was conducted in the major academic databases IEEE Xplore[1], SpringerLink[2], and ScienceDirect[3] - focusing on publications from 2021 to 2025. The search strategy used specific keywords related to AI and SE (e.g. 'Artificial intelligence' AND 'social engineering', 'deepfake' AND 'cybersecurity'). The initial screening identified 38 relevant documents. Applying topic relevance, empirical rigor, and focus criteria to the intersection of AI and SE, the final sample was narrowed to 12 high-quality publications for in-depth analysis.

To provide a clear overview of the selected articles before delving into thematic and analytical discussions in subsequent sections, a summary table 1 is presented below, which synthesizes key characteristics of each article, including publication year, AI technologies employed, attack types addressed, main findings, and proposed mitigation strategies.

## 3 Technological Foundations of AI in SE

### 3.1 Historical and Current Applications of ML in SE

Although the use of AI in SE dates back to early tools like *CyberLover* (2007), a chatbot that uses Natural Language Processing (NLP) to mimic conversations and extract personal data, its impact has grown significantly with ML advancements [4]. The rise of social media bots that utilize computer scripts to emulate human behaviors and manipulate discussions marked a shift toward "spreading misinformation, opinion manipulation, and coordinated inauthentic campaigns" [12]. For example, a sophisticated tool like *DeepLocker*, a malware that uses a Deep Neural Network (DNN) AI model to hide its attack payload and activates only when the intended target is reached, demonstrates the precision of modern AI driven attacks [12].

Currently, recent advances in machine learning have increased the sophistication of SE attacks, which leverage supervised and unsupervised techniques to increase the capabilities, effectiveness, and scalability of malicious attacks. These models can generate convincing phishing emails and fake social media profiles by identifying

[1] `https://ieeexplore.ieee.org/Xplore/home.jsp`
[2] `https://link.springer.com`
[3] `https://www.sciencedirect.com`

**Table 1.** Summary of Selected Articles from Systematic Review

| Citation | Year | AI Technologies Used | Attack Types Analyzed | Key Findings or Conclusions | Mitigation Strategies Proposed |
|---|---|---|---|---|---|
| [1] | 2024 | Generative AI, Chatbots, Deepfakes | Phishing, Smishing, SE (gereral) | AI-generated content automates and enhances realism and personalization in SE attacks, increasing success rates. | Detection via persuasion-aware classifiers; user awareness campaigns tailored to different professional groups. |
| [2] | 2024 | Various AI techniques | SE, Misinformation/Fake News, Hacking | AI advancements increase cybersecurity risks, enabling sophisticated SE, disinformation, and hacking. | Emphasis on transparency, explainability, and ethical AI use; translation of technical concepts for broader audiences. |
| [3] | 2025 | Deepfake (voice, facial synthesis) | Phishing, Business Email Compromise (BEC), Identity Theft | Deepfakes enhance SE by creating ultra-realistic audiovisual content that exploits human trust. | Regulatory oversight; detection systems; public education. |
| [4] | 2022 | General AI, Machine Learning (ML) | Integrity attacks, Misinformation, SE, Autonomous systems | Classifies malicious uses of AI, highlighting expansion of vulnerabilities and new threats. | Governance mechanisms; interdisciplinary responses; ongoing threat assessment. |
| [6] | 2023 | ChatGPT, FraudGPT, WormGPT, DALL-E 2, Stable Diffusion, VALL-E | Phishing, Pretexting, Deepfake SE, Voice Cloning | LLMs enable personalized SE attacks at scale, using manipulation and synthetic content. | AI-based detection tools; Multi-Factor Authentication (MFA); phishing simulations; zero trust; continuous updates. |
| [7] | 2025 | Deepfake, Generative Adversarial Networks (GANs) | Misinformation, Fraud, Political Manipulation | Deepfakes are powerful tools of deception, undermining trust in digital media. | Criminalization of malicious use; AI watermarking; public awareness; multi-stakeholder governance. |
| [8] | 2024 | LLMs, ML, AI malware | Phishing, CAPTCHA bypass, Voice Cloning, Ransomware | AI tools automate and enhance malicious activities, making attacks faster and harder to detect. | ML-based anomaly detection; employee training; MFA; verbal verification techniques. |
| [9] | 2025 | ML classifiers, Artificial Neural Networks (ANN), Blockchain (IPFS, Smart Contracts) | Phishing, Smishing, Vishing, Piggybacking, Quid Pro Quo | AI-enabled SE is widespread; hybrid systems improve detection and data integrity. | AI for pattern recognition; blockchain for URL storage; MFA; encryption; behavior monitoring. |
| [10] | 2024 | Generative AI (LLMs, GANs) | Phishing, Impersonation, Vishing | GenAI enhances SE through realism, automation, and personalization. | AI-based deception detection; watermarking; privacy tools; media literacy; rate limiting. |
| [11] | 2023 | ChatGPT, Bard, Llama2, Claude; Prompt Engineering | Persuasion-based SE | LLMs can be manipulated via subtle prompts to bypass protections. | Countermeasures integrating psychology and technical safeguards; human-in-the-loop validation. |
| [12] | 2024 | LLMs, Diffusion Models, Reinforcement Learning | Multi-phase SE (3E: Enlarging, Enriching, Emerging) | Proposes 3E model for AI-SE evolution; calls for robust risk analysis. | Markov Decision Process (MDP)-based risk quantification; adaptive defenses; content filtering; anonymization. |
| [5] | 2020 | AI/ML, Reinforcement Learning (RL), GANs, Neural Nets | Spam, Malware, Phishing, Voice Cloning | AI improves traditional attacks and lowers entry barriers via AI-as-a-Service. | Early detection; cloud monitoring; CAPTCHA filter detection; poisoning dataset defense. |

patterns in legitimate and fraudulent data, such as linguistic cues, sender metadata, or account behaviors like posting frequency and writing style [10] [1]. Furthermore, their "analysis can identify potential targets, assess their vulnerabilities, and predict their behavior based on patterns" [10], amplifying the effectiveness and reach of AI-driven SE operations.

As highlighted by [10], "attackers can leverage AI algorithms to analyze large datasets and create highly targeted phishing campaigns on-scale. This automation and personalization significantly increase the effectiveness of spear phishing campaigns, posing a greater threat to individuals and organizations".

### 3.2 Large Language Models and Prompt Exploitation

Another critical facet of AI-driven SE is the manipulation of large language models LLM through prompt engineering. Models such as ChatGPT[4], Claude[5] and Gemini[6] demonstrated vulnerability to prompt injection attacks, particularly "Jail-breaking Prompt" [11]. These attacks "can manipulate the model's responses and bypass content filters, raising ethical and security concerns" [11]. Research shows that prompts grounded in psychological principles such as authority, scarcity, urgency, and reciprocity can override ethical safeguards and elicit harmful results from LLMs [11]. Furthermore, prompt injection attacks exploit ambiguous or hidden instructions within inputs to bypass model constraints, allowing the creation of disinformation, impersonation content, or unintended behaviors [12] [10].

The LLMs specifically designed for offensive purposes, such as FraudGPT [7], promises to amplify the sophistication of these attacks [11]. As the authors report, these models are promoted as versatile tools that operate without limitations or boundaries. They are distributed on the dark Web and include features designed for criminal operations, such as malware generation, phishing email automation, and unrestricted content formatting [11] [6]. "These tools democratize cybercrime by lowering the technical barrier for malicious actors" note [6].

Another purpose for which LLMs can be used by criminals is to integrate it into interactive chatbots that allow attackers to conduct real-time conversations with victims, increasing engagement and reducing suspicion [10]. Additionally, these bots can be distributed on a large scale and repeatedly, which is further amplified when considering their success rate. This ability represents a significant shift from traditional static phishing, as it "can supercharge deceptive campaigns, making them highly sophisticated and more challenging to identify and counter" [10].

### 3.3 The Rise of Generative AI and Synthetic Media: Deepfakes and Voice Cloning

Generative AI - a subfield of AI focused on producing synthetic content with characteristics similar to humans - has significantly amplified the capabilities of SE. Its algorithms can generate text, images, audio and video that are nearly

[4] `https://chat.openai.com`
[5] `https://claude.ai`
[6] `https://deepmind.google/technologies/gemini`
[7] See: `https://netenrich.com/blog/fraudgpt-the-villain-avatar-of-chatgpt`

indistinguishable from real media, fueling the rise of synthetic content such as deepfakes - videos produced using GANs - and voice cloning [6].

Deepfakes, for instance, have achieved a level of realism capable of reproducing facial expressions, lip movements, and gestures of known individuals with extreme precision. GANs operate through two competing neural modules: a generator, which creates fake images, and a discriminator, which attempts to distinguish synthetic from real content, resulting in a constant refinement cycle [5] [12] [3]. Platforms like DALL · E 2[8], Stable Diffusion[9], and Synthesia[10] further extend these capabilities, allowing the generation of convincing fake images and unique digital faces, often used in the creation of fictitious social media profiles, romance scams, and corporate fraud schemes [6]. The accessibility of these tools, including mobile apps, has facilitated their use by malicious actors with little or no technical experience [12]. For example, [10] highlights a case in which an employee was deceived into transferring $25 million after deepfake technology was used to mimic the voice and appearance of the chief financial officer. Interestingly, all the other people present at the video conference were also deepfakes.

In parallel, voice cloning uses DNNs trained on short audio samples to replicate a person's tone, rhythm and inflection. Advanced tools like Microsoft's Neural Codec Language Model for Text-to-Speech (VALL-E)[11] can generate convincing voice simulations from just a few seconds of input, enabling attacks such as business fraud and audio-based virtual kidnappings [6]. Another notable case involved a UK-based energy firm that lost £220,000 after a worker, fooled by a cloned voice impersonating the CEO of its parent company, authorized a fraudulent fund transfer demonstrating how voice synthesis can bypass trust-based verification mechanisms [1].

As can be seen, the growing sophistication of these tools presents an urgent challenge in developing technical, educational, and regulatory mechanisms that ensure the authenticity of the content, the verification of the source, and the protection against manipulation.

### 3.4 Real-world Illustration

As a real-world example of the power of AI in social engineering attacks, we highlight the case reported by Schmitt and Fléchais [10], which occurred in 2024. In this incident, cybercriminals used advanced deepfake technology to simulate a videoconference involving the CFO of a multinational company. The attackers created synthetic personas of several executives by cloning their voices and generating deepfake video avatars. In a seemingly authentic and authoritative virtual meeting, they exploited classic social engineering principles such as urgency and hierarchical trust to convince an employee to authorize a $25 million bank transfer [12]. This case demonstrates that generative artificial intelligence is no longer a hypothetical threat; it has become an operational tool for large-scale deception, capable of bypassing traditional verification mechanisms in high-stakes environments.

[8] `https://openai.com/dall-e-2`
[9] `https://stability.ai/stable-diffusion`
[10] `https://www.synthesia.io`
[11] `https://www.microsoft.com/en-us/research/project/vall-e-x/vall-e/`
[12] https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html

## 4 Modeling AI-Driven SE: Comparative Frameworks

The following subsections explore three different frameworks proposed in the recent literature, highlighting their structure, methodology, and contributions to the ongoing discourse on digital deception.

### 4.1 Generative AI Social Engineering Framework

The Generative AI Social Engineering Framework (GenAI-SE), proposed by [10], is a multidimensional technical model designed to analyze the influence of generative artificial intelligence capabilities on SE attacks. Structured around three axes, it enables the decomposition and evaluation of how AI increases the effectiveness and reduces the operational cost of these attacks at scale.

In the dimension of realistic content generation, the authors emphasize the use of deepfakes and other synthetic media, which significantly increase the difficulty of detection. The advanced targeting and personalization layer highlights the use of large datasets (e.g., social media, public leaks) to tailor attacks. Finally, the automated attack infrastructure layer, how AI services can accelerate large-scale deployment and power social engineering attacks. Services such as Software as a Service (SaaS) platforms and the gig economy[13] enable non-experts to easily access, configure, and execute AI-powered attacks without requiring advanced technical skills.

The framework also presents five analytical dimensions of AI-enabled social engineering attacks, which guide its practical application in the cybersecurity context. The first dimension, SE Attack Lifecycle allows examination of all stages in the life cycle of an attack; the second, SE Attack Type, assesses the impact of AI on different types of attacks; the third, Generative AI Models compares the capabilities and weaknesses of AI models in the context of SE; the Countermeasures dimension supports the identification and development of defensive strategies; and finally, Implications address the social, legal, and ethical consequences of integrating AI into SE.

To measure these effects, the authors introduce two core indicators: Threat Amplification, referring to the increased reach and success of attacks, and Cost-Effectiveness, concerning the reduced operational costs. They argue that AI enhances not only the effectiveness but also the accessibility of attacks, calling into question the adequacy of awareness-based defenses. Recommended countermeasures span technical (e.g., digital watermarking, AI-based detection) and human-centric approaches (e.g., media literacy, regulatory measures).

### 4.2 Generative AI Social Engineering Framework

The study conducted by [1] adopts a user-centered analytical approach that focuses on the experiences of victims with AI-generated SE content. Data were collected through semi-structured interviews with 40 victims SE and chatbot usability tests

[13] The gig economy is a labor market characterized by the prevalence of short-term contracts or freelance work. See: Sundararajan, A. (2016). The gig economy is coming. Harvard Business Review. `https://hbr.org/2016/02/a-sharing-economy-needs-a-sharing-government`

with 40 additional participants. The results highlight that emotional personalization and detection difficulty were decisive in the success of the attack. Recommendations include AI-driven defenses, continuous user education, and mitigation strategies aligned with user behaviors and vulnerabilities.

### 4.3 The 3E Phases Framework and AI Risk Quantification Model

The study conducted by [12] proposes a categorization that evolves the nature of AI-powered SE attacks. This model is structured into three phases: (i) Enlarging: Scaling traditional attacks through automation (e.g., mass phishing, data scraping); (ii) Enriching: Personalizing messages using deep-learning techniques; and (iii) Emerging: Empowering attacks by LLMs, enabling adaptive real-time strategies and large-scale campaigns.

Furthermore, the authors propose a quantitative risk assessment framework for AI-powered SE attacks, based on MDP, "as a foundation for comparative impact evaluations, forensic traceability of sociotechnical factors, and prioritization of mitigation measures". According the authors "The framework could augment SE attack propagation dynamics and causally influence downstream social impacts over time. This involves leveraging big data to track changes in key risk metrics like spreading capability and penetration efficiency under different conditions".

### 4.4 Analytical Discussion: Impacts of AI on Social Engineering

The joint analysis of the three studies reveals a significant methodological convergence in how AI is integrated into SE attack models. Despite formal differences in the proposed frameworks, we can observe that attacks powered by AI can be decomposed into three fundamental elements: realistic content generation, contextual message personalization, and automated or scalable attack execution.

These models highlight a consolidated trend in the literature: SE attacks mediated by AI are frequently understood as complex systems that combine technical and human elements, which means that their effectiveness is based not only on technology, but on the balance between persuasive language, personalized targeting, and the ability to coordinate large-scale attacks. Each study confirms that advances in AI not only increase the manipulative power of malicious content, but also reduce technical and operational barriers for attackers at different skill levels. The boundary between human interaction and automated action is increasingly blurred, increasing the overall complexity of defensive strategies.

Moreover, all three studies underscore that traditional mitigation approaches – such as user training and signature-based filtering – are insufficient in the face of highly personalized and contextually relevant attacks. The need for integrated countermeasures that combine behavior-based automated detection, restrictions on personal data exposure, and explainable algorithmic defenses is consistently emphasized in the reviewed literature. Finally, the impact of AI extends beyond technical considerations, touching on ethical, cognitive, and regulatory domains. It demands a repositioning of cybersecurity policies, with particular emphasis on developer accountability, AI model governance, and algorithmic transparency.

## 5 Challenges and Ethical Implications

The growing use of AI in digital contexts also brings to light ethical and legal concerns. Recent literature identifies that these risks come from the opacity of AI systems - which often operate as "black boxes", making it difficult to understand how they function and how decisions are made — and the lack of accountability, the large-scale capability to manipulate human perception, and the difficulty in detecting and defending against SE attacks instrumentalized by AI [5]. This issue is emphasized by [4], who identify algorithmic opacity as a first-order ethical challenge, as it obstructs audits of system errors, biases, or embedded abuse. According to the authors, this lack of transparency can be deliberately exploited by malicious actors who manipulate decision-making systems or embed harmful instructions in ways that are practically undetectable.

The ethics of authenticity is also deeply affected by deepfakes, which damage the reputations of individuals and institutions. The situation is further exacerbated by the high false-negative rates of current deepfake detection systems, fueling a vicious cycle in which the malicious use of AI advances faster than available countermeasures. Another critical concern involves the misuse of personal data for offensive purposes. As noted in [2], AI algorithms have been used to collect and process large amounts of data without the knowledge or consent of individuals, to build highly accurate psychosocial profiles.

In conclusion, it is observed that the ethical challenges associated with AI are based on its facilitation of opaque, abusive, and unaccountable practices. What emerges is a digital ecosystem in which automated decisions, hyperrealistic simulations, and hidden inferences become a frequent part of everyday life, undermining fundamental guarantees such as privacy, information authenticity, and individual autonomy. Addressing these challenges requires more than technical fixes; it requires the development of a set of regulatory measures and best practices capable of responding to the complex moral and legal dilemmas posed by AI in high-risk environments.

## 6 Conclusion

Generative AI has significantly amplified the sophistication, scale, and accessibility of SE campaigns. Our review of recent studies reveals that AI-mediated attacks enhance effectiveness and reach while reducing operational costs, democratizing these tactics for less skilled actors.

The case studies analyzed converge on key pillars that allow highly persuasive and customized messages to be distributed on a large scale, using technologies such as LLMs, GANs and neural voice cloning. As observed, emotional personalization and communicative realism are critical success factors explored by AI, with overall effectiveness further amplified by a substantial decrease in execution costs. However, despite the technical advances in countermeasures, the authors unanimously caution that a purely technical approach is insufficient. The ethical and regulatory challenges require a reorientation of cybersecurity policies towards responsible governance models with accountability, transparency, and social oversight.

Our review identifies critical research gaps that underscore the urgent need for further investigation. There is a lack of empirical validation across diverse sociocultural and occupational groups, which limits the broader applicability of the current findings. Integrated analyses of the computational, psychological and linguistic dimensions of AI-generated suggestion mechanisms are scarce, hindering our understanding of vulnerability factors. Concrete proposals for automated detection of synthetic persuasion patterns, especially in real-time, are also lacking. Finally, stronger connections between technical measures and normative frameworks are needed to inform ethical AI governance in offensive applications.

Given these points, it can be considered that addressing the emerging threats posed by AI-based SE requires an interdisciplinary response that combines technological innovation, a deep understanding of human factors, and ethical foresight. Only through this integration will it be possible to develop defense strategies that are not only effective but also fair and sustainable in the face of increasingly sophisticated digital threats.

## References


1. Alahmed, Y., Abadla, R., Al Ansari, M.J.: Exploring the potential implications of ai-generated content in social engineering attacks. In: 2024 International Conference on Multimedia Computing, Networking and Applications. IEEE (2024)
2. Anwar, S., Perez, A.: Risks and ethical concerns in cyber security with advancements of artificial intelligence – a systematic review. Tech. rep., Texas A&M University (2024)
3. Blake, H.: Ai-powered social engineering: Understanding the role of deepfake technology in exploiting human trust (2025), disponível em: ResearchGate. Acesso em: 10 abr. 2025
4. Blauth, T.F., Gstrein, O.J., Zwitter, A.: Artificial intelligence crime: An overview of malicious use and abuse of ai. IEEE Access **10**, 77110–77122 (2022)
5. Ciancaglini, V., Gibson, C., Sancho, D., McCarthy, O., Eira, M., Amann, P., Klayn, A.: Malicious uses and abuses of artificial intelligence. Tech. rep., Trend Micro Research; United Nations Interregional Crime and Justice Research Institute (UNICRI); Europol's European Cybercrime Centre (EC3) (2020), `https://www.europol.europa.eu`, accessed: 2025-04-23
6. Falade, B.: Fraudgpt and wormgpt: A growing threat. Cybersecurity Monthly (jul 2023)
7. Folorunsho, F., Boamah, B.F.: Deepfake technology and its impact: Ethical considerations, societal disruptions, and security threats in ai-generated media. Tech. rep., Ball State University (2025)
8. Kamruzzaman, A.S., Thakur, K., Mahbub, S.: Ai tools building cybercrime & defenses. In: International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA 2024). IEEE (2024)
9. Rathod, T., et al.: A comprehensive survey on social engineering attacks, countermeasures, case study, and research challenges. Information Processing & Management **62** (2025). https://doi.org/https://doi.org/10.1016/j.ipm.2024.103928
10. Schmitt, M., Fléchais, I.: Digital deception: Generative artificial intelligence in social engineering and phishing. Artificial Intelligence Review **57**, 324 (2024). https://doi.org/https://doi.org/10.1007/s10462-024-10973-2
11. Singh, A., et al.: Exploiting large language models (llms) through deception techniques and persuasion principles (2023), disponível em: https://arxiv.org/abs/2309.02244. Acesso em: 16 abr. 2025
12. Yu, J., et al.: The shadow of fraud: The emerging danger of ai-powered social engineering and its possible cure. arXiv preprint arXiv:2407.15912 (2024)

# Bias & Accountability of Artificial Intelligence in Cancer Diagnosis: A Conceptual Reflection

Nelson Faria[1], Sofia Campelos[2], and Vítor Carvalho[1]

[1] 2Ai – School of Technology, IPCA, Barcelos, Portugal
a14805@alunos.ipca.pt,vcarvalho@ipca.pt
[2] IMP Diagnostics, Porto, Portugal
sofia.campelos@gmail.com

**Abstract.** Bias and accountability are critical ethical challenges in the application of artificial intelligence (AI) for cancer diagnosis. AI-powered computer-aided diagnosis systems have the potential to enhance diagnostic accuracy and reduce clinician workload, but their effectiveness is contingent on how well they address these ethical concerns. Bias in AI models, often caused by unrepresentative data, can lead to misdiagnoses that disproportionately harm certain groups, deepening existing health inequalities. Considering the literature, we believe that it's not enough to simply create these technologies but also to think about their risks and the people they will affect. In academia, the rush for innovation can sometimes push transparency and thorough testing to the back burner. On the political side, regulations often come too late, and funding priorities favour speed over ethical responsibility, leaving a gap in accountability. Socially, when AI systems fail and no one takes responsibility, it weakens public trust and can worsen disparities in healthcare. Addressing both bias and accountability requires inclusive data practices, transparent research, and clear governance, ensuring AI systems serve all patients equitably and safely.

**Keywords:** Accountability · Artificial Intelligence · Bias · Cancer Diagnosis · Ethics.

## 1 Introduction

Artificial Intelligence (AI) is increasingly demonstrating its capabilities across various domains, including healthcare [4]. Its ability to process vast amounts of data, recognize complex patterns, and assist in decision-making has made it a valuable tool in medical diagnostics. In cancer diagnosis, AI-powered computer-aided diagnosis (CAD) systems have the potential to improve diagnostic accuracy and assist clinicians in identifying tumour tissues and reducing their workload [5, 18, 20].

As AI continues to evolve in cancer diagnosis, academic institutions have been crucial in developing and integrating AI models into clinical practice [9]. Beyond building theoretical models, their role also includes rigorous testing, validation, and refinement in order to ensure that the developed AI models are safe and suitable for clinical use [9,24]. The cooperation between researchers, clinicians, and regulatory bodies facilitates the implementation of theoretical advancements into practical diagnostic systems that align with clinical workflows and patient needs, while adhering to ethical and methodological standards [12, 14, 21].

Moreover, academia is committed to thorough and unbiased research that prioritizes patient welfare, scientific integrity, and long-term societal impact [8]. Although external pressures, such as the imperative to publish and secure funding, may affect these ideals, it remains essential that academic research follows these

standards [10]. Therefore, scrutinizing newly developed AI models through peer-reviewed publications and academic conferences is a key responsibility [10, 13]. These venues are examples of opportunities to evaluate the quality, scale, and representativeness of the used datasets, and to verify that the reported results are meaningful, trustworthy, and reproducible [1, 16]. This process of rigorous review addresses concerns such as fairness, transparency, and effectiveness before clinical implementation [16]. Through this oversight, academia strengthens the reliability of AI tools in cancer diagnosis and contributes to responsible innovation.

This paper aims to identify and critically reflect on the ethical challenges surrounding AI-powered cancer diagnosis systems. The reflection will address how these issues intersect with personal, academic, political, and social lenses, emphasising the responsibilities of researchers, institutions, and regulatory bodies in ensuring that AI tools are developed and implemented in ways that are equitable, transparent, and safe.

## 2 Ethical Issues of Artificial Intelligence in Cancer Diagnosis

The continuous integration of AI models into cancer diagnosis arose ethical issues, particularly concerning bias and accountability [6]. Since these models will be applied in healthcare systems, it is essential to examine their behaviour when analysing patterns across diverse patient populations and how the responsibility is distributed when an error occurs [1, 6].

The models' bias generally comes from training them with datasets that lack diversity in terms of ethnicity, gender, age, or disease representation, leading to models that may perform satisfactorily on average but poorly when presented with data of minority groups or underrepresented disease subtypes [6]. In cancer diagnosis, errors like false positives or negatives and misclassification of the cancer subtype significantly impact the well-being and quality of life of the patients, making them undergo unnecessary treatments or delaying the correct diagnosis and the postponement of necessary medical interventions [5, 19]. [3, 7, 19] are some examples of studies that provide evidence of the bias that AI models can suffer when trained with datasets that do not represent the full spectrum of patients. In [19], the authors demonstrated that AI algorithms applied to chest radiographs exhibit underdiagnosis biases, particularly affecting intersectional underserved subpopulations. The authors of [3] evaluated the generalizability of a deep learning system for screening mammography on a local dataset and found that they could not fully replicate the algorithm's originally reported performance. More recently, in [7], Huang et al. investigated fairness in deep learning models for breast cancer diagnosis and found biases linked directly to the racial and ethnic composition of training datasets. Collectively, these studies indicate that bias is a prevalent challenge in AI-driven cancer diagnosis and underscore the urgent need for improvements in dataset diversity and model validation.

At the same time, the increasing complexity and autonomy of AI-powered CAD systems raise significant accountability concerns. When these tools contribute to diagnostic errors, the consequences directly affect treatment plans and patient

outcomes. However, determining who is responsible in such cases remains unclear. The fault may lie with the developers, researchers, clinicians, or institutions that deployed the systems. This ambiguity can hinder proper accountability and undermine public confidence in the safety and reliability of AI in healthcare [2, 6, 17]. For instance, [2, 17] demonstrate that the multifaceted nature of AI systems complicates the assignment of liability because the origin of the errors may come from different stages. Similarly, [14] highlights that current accountability frameworks struggle to keep pace with the autonomous capabilities of modern AI, leaving significant gaps in determining responsibility.

## 3 Critical Reflection on Bias in Artificial Intelligence for Cancer Diagnosis

### 3.1 Personal Lens

Bias is a complex topic to address, especially in the healthcare domain. In the context of cancer diagnosis, biased outcomes represent a serious concern, as diagnostic errors can profoundly affect a patient's trajectory, both medically and emotionally [5]. Bias is further exacerbated when AI-powered CAD systems are involved, whose advertised objectivity and increasing influence on clinical decision-making may obscure underlying errors. To mitigate this, AI models used in these systems must be as unbiased as possible, and achieving this requires diverse and representative datasets that include a wide range of variables, such as age, gender, ethnicity, and cancer subtypes. Having a model trained on a dataset composed of the full spectrum of cases enables it to acquire better knowledge of clinical variability and produce more accurate outcomes. When this level of inclusivity is not met, it can significantly impact a person's life. In the case of a false positive or misdiagnosis of the cancer subtype, the patient will undergo unnecessary invasive treatments, face avoidable health expenses, and may take the place of someone who really needs it. On the other hand, if a diagnosis is missed or delayed, cancer will progress untreated and may even turn a potentially manageable condition into a fatal one, causing the loss of a human life. Additionally, the model can be considered discriminatory due to the exclusion of segments of the population. This not only exacerbates existing inequalities in healthcare access and outcomes but also can lead to a global loss of confidence in AI-powered CAD systems, making people question whether these systems should be accepted and call into doubt the credibility of scientific research and technological innovation, leading to feelings of frustration and despair. In more critical scenarios, a lack of trust can threaten the adoption and continuity of these systems, resulting in wasted resources, loss of prior research efforts, and job displacement.

As researchers in this field, the challenge of bias in AI-powered CAD systems directly impacts our values and actions. Transparency, fairness, effectiveness, and dignity are some of the key principles that we consider essential for researchers when building these AI models, especially since these systems are being designed to support decisions that can extend or preserve human life. With this in mind, the worth of a human life is paramount, and every design choice should reflect that priority. The development of these models shouldn't only focus on being faster and more innovative but also on

serving a wide range of patients reliably. At the same time, we acknowledge that it may be difficult, if not impossible, to cover the full spectrum of human variability in training datasets, mainly due to legal, ethical, and logistical constraints. Still, these adversities do not diminish the researchers' responsibility to strive for equitable representation and anticipate the risks of exclusion. This search for inclusiveness comes with trade-offs and risks, such as slowing down progress and falling behind competitors, creating friction in fast-paced innovative environments, or going against institutional pressures focused on rapid publication or commercialization. Nonetheless, the integrity of research and respect for the value of human lives must take precedence over these concerns. As a result, thinking about this issue has deepened our sense of responsibility, and addressing bias is not just a technical objective but a moral imperative, one that compels us to advocate for inclusive data practices and greater transparency in reporting model limitations.

### 3.2 Academic Lens

Academia can be seen as both a contributor to the problem of bias in AI and a potential place for its resolution. The constant pressure to publish quickly, secure funding, and produce innovative and marketable technologies often leads researchers to prioritize faster development and novelty over inclusivity and ethical rigour [10]. While these choices are rarely made with malicious intent, they end up reinforcing a culture where addressing bias is deprioritised. Additionally, systemic barriers such as limited access to representative datasets and a lack of interdisciplinary collaboration can significantly shape the direction of the research being conducted [6, 22]. These constraints can limit researchers' ability to fully understand how their work connects to its real-world impact on human lives. At the same time, academia frequently sets the research agenda and influences which questions are asked and funded. For these reasons, academia is the right space to meaningfully change processes, promoting equitable AI, encouraging critical reflection, and emphasizing the importance of fairness and transparency in scientific reporting [2, 6, 14].

### 3.3 Political Lens

One of the main sources of research funding in academia is government or politically influenced institutions [15]. These entities often shape the direction and priorities of research projects, favouring those that promise innovation, field leadership, and economic return [6, 15]. However, these criteria do not always align with most ethical principles and can perpetuate bias in AI models. Furthermore, policy-making and regulatory bodies normally set ethical boundaries only after technological advancements have already been made. This delay allows researchers to progress more "freely" without adequate consideration of ethical implications, potentially worsening health outcomes of already underserved patients as biased tools are adopted prematurely [1, 23]. Thus, political institutions and regulatory interventions are essential to valuing fairness over innovation.

### 3.4 Social Lens

The premature adoption of AI-powered CAD systems can have societal consequences, particularly when their AI models are trained on datasets that underrepresent certain populations or cancer subtypes. Patients belonging to these groups face a greater risk of having a false outcome or a delayed diagnosis. As already mentioned, the health disparities will exacerbate, and this will erode public trust in these systems. When systems consistently fail minority groups, both patients and clinicians will be sceptical about adopting other scientific advances that claim to be objective but do not equitably serve everyone. However, resolving these issues requires engagement with affected communities, transparency about model limitations, and the application of techniques like synthetic data generation or fairness-aware augmentation to develop a broader range of cancer subtype cases in scenarios where representative data is unavailable [1, 6, 14].

### 3.5 Summary

Addressing bias in AI-powered CAD systems is not just a technical challenge but also an ethical responsibility. From a personal lens, we believe that the value of human life must always take precedence over the pressures of innovation. The consequences of biased outcomes can deeply affect patients and entire communities, exacerbating existing disparities in healthcare. While achieving fully representative datasets may be difficult, attempting to obtain inclusiveness in AI development is essential, even if it means slowing down progress or confronting institutional pressures. The principles of research integrity, transparency regarding model limitations, and a commitment to fairness must guide the development of these systems to ensure they serve all patients equitably, regardless of their characteristics, background, or circumstances.

## 4 Critical Reflection on Accountability in Artificial Intelligence for Cancer Diagnosis

### 4.1 Personal Lens

The AI models in CAD systems are extremely useful to help clinicians with cancer diagnosis, improving the accuracy of the diagnosis and reducing the workload [5, 20]. However, when these systems fail by giving a false outcome, it might influence whether a person gets a life-saving diagnosis or not, and blame can shift between developers, researchers, clinicians, or institutions [2, 17]. From our perspective, we do not think accountability should be a game of deflection. If we are contributing to building these tools, then we should also be thinking about how they will be used, what their risks are, and who could be affected. Developers must clearly document model behaviour and limitations; clinicians need to consider AI-powered CAD systems as an assistant instead of relying 100% on them and have protocols for double-checking AI output; institutions should establish incident-response procedures; and researchers must design systems with built-in audit trails and error-logging. From our point of view, only by clearly defining roles, maintaining transparency of decision

logs, and anticipating failures can we guarantee accountability, security, and that these technologies truly serve the patients who rely on them.

### 4.2 Academic Lens

Research institutions often lead technological advancements, measuring their success through metrics such as novelty, publication volume, and grant acquisition. Yet, long-term safety and responsibility for their downstream effects are factors that tend to be poorly defined or deprioritised [9, 24]. Publications regarding AI models for cancer diagnosis are no exception and they may be published and promoted with limited transparency about limitations, testing conditions, or intended use [6]. Therefore, when errors occur later in clinical practice, the accountability often shifts to clinicians. This attitude challenges the credibility of academic research and raises concerns about research integrity [17]. To address this gap, academia must look beyond innovation and take responsibility for how their AI models are developed, trained, validated, and distributed, including offering support in case of failure in their algorithms. Explainable AI tools can support accountability by making model predictions and reasoning processes more transparent to clinicians, enabling better error detection and model trustworthiness.

### 4.3 Political Lens

Additionally, AI-powered CAD systems may be integrated into clinical workflows before adequate governance frameworks are in place, exposing patients to unvetted risks and with no clear policies outlining who is ultimately accountable. Moreover, there is a tendency where regulatory bodies only act after failures occur and funding institutions prioritize innovation and economic growth over ethical standards. These factors contribute to a diffusion of responsibility, which leaves patients and frontline clinicians bearing the risks. As a result, there is an urgent need to implement clear accountability mandates, promote patient-centred regulations, and fund research that values ethical oversight as much as performance [17].

### 4.4 Social Lens

From a social perspective, accountability is an important factor in protecting patients when an AI model fails in a cancer diagnosis. Underrepresented groups may be more likely to be misdiagnosed or excluded from training data, leading to diagnostic errors that go unrecognized and unaddressed. These failures reinforce existing inequalities in healthcare access, quality, and outcomes [1, 6, 21]. Furthermore, ensuring that someone will take responsibility in case of a misdiagnosis improves public trust in medical technologies. When patients know that someone is responsible and that there are processes to respond to failures, they are more likely to feel safe using and accepting AI-supported care [2, 6, 11, 21].

### 4.5 Summary

Accountability is an important ethical factor in AI-driven cancer diagnosis. When these systems make mistakes, someone must be clearly responsible to protect patients and maintain trust. This means defining roles, creating regulations, and ensuring transparency throughout the research and deployment process. Without clear accountability, the risks of harm, inequality, and public mistrust grow. If no one is held accountable when AI fails, then the systems meant to save lives may end up putting them at risk.

## 5 Final Remarks

AI holds transformative potential for cancer diagnosis, but its integration into healthcare systems raises significant ethical concerns, particularly regarding bias and accountability. This study has reflected how these issues intersect with different lenses, such as personal values, academic responsibilities, political institutions, and social consequences. While AI can support earlier and more accurate diagnosis, these benefits are not guaranteed for all populations unless equity, transparency, and responsibility are prioritized throughout the development of the AI models.

Academia acts both as a contributor to ethical risks when pressured by innovation-driven incentives, and as a key site to promote critical scrutiny, fair data practices, and interdisciplinary collaboration. Politically and socially, delayed regulation, innovation-centric funding priorities, biased outcomes, and the lack of clear accountability allow gaps in oversight that can harm underrepresented patient groups, deepen health disparities, and erode public trust in medical technologies.

To ensure that AI tools genuinely serve all patients, ethical principles must be embedded not only in the models but in the institutions and systems that create and govern them. The future of AI in cancer diagnosis will be shaped not just by technical breakthroughs but by the willingness of all stakeholders to take responsibility for its impact.

## References


1. Albahri, A.S., Duhaim, A.M., Fadhel, M.A., Alnoor, A., Baqer, N.S., Alzubaidi, L., Albahri, O.S., Alamoodi, A.H., Bai, J., Salhi, A., Santamaría, J., Ouyang, C., Gupta, A., Gu, Y., Deveci, M.: A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. Information Fusion **96**, 156–191 (Aug 2023). https://doi.org/10.1016/j.inffus.2023.03.008, https://www.sciencedirect.com/science/article/pii/S1566253523000891
2. Amann, J., Blasimme, A., Vayena, E., Frey, D., Madai, V.I., the Precise4Q consortium: Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Medical Informatics and Decision Making **20**(1), 310 (Nov 2020). https://doi.org/10.1186/s12911-020-01332-6, https://doi.org/10.1186/s12911-020-01332-6
3. Condon, J.J.J., Oakden-Rayner, L., Hall, K.A., Reintals, M., Holmes, A., Carneiro, G., Palmer, L.J.: Replication of an open-access deep learning system for screening mammography: Reduced performance mitigated by retraining on local data (Jun 2021). https://doi.org/10.1101/2021.05.28.21257892, https://www.medrxiv.org/content/10.1101/2021.05.28.21257892v1, iSSN: 2125-7892 Pages: 2021.05.28.21257892

4. Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., Dean, J.: A guide to deep learning in healthcare. Nature Medicine **25**(1), 24–29 (Jan 2019). https://doi.org/10.1038/s41591-018-0316-z, https://www.nature.com/articles/s41591-018-0316-z, publisher: Nature Publishing Group
5. Faria, N., Campelos, S., Carvalho, V.: A Novel Convolutional Neural Network Algorithm for Histopathological Lung Cancer Detection. Applied Sciences **13**(11), 6571 (Jan 2023). https://doi.org/10.3390/app13116571, https://www.mdpi.com/2076-3417/13/11/6571, number: 11 Publisher: Multidisciplinary Digital Publishing Institute
6. Hanna, M.G., Pantanowitz, L., Jackson, B., Palmer, O., Visweswaran, S., Pantanowitz, J., Deebajah, M., Rashidi, H.H.: Ethical and Bias Considerations in Artificial Intelligence/Machine Learning. Modern Pathology **38**(3), 100686 (Mar 2025). https://doi.org/10.1016/j.modpat.2024.100686, https://www.sciencedirect.com/science/article/pii/S0893395224002667
7. Huang, K., Wang, Y., Xu, M.: Investigating the Fairness of Deep Learning Models in Breast Cancer Diagnosis Based on Race and Ethnicity. Proceedings of the AAAI Symposium Series **4**(1), 303–307 (Nov 2024). https://doi.org/10.1609/aaaiss.v4i1.31806, https://ojs.aaai.org/index.php/AAAI-SS/article/view/31806
8. Jackson, B.R., Ye, Y., Crawford, J.M., Becich, M.J., Roy, S., Botkin, J.R., de Baca, M.E., Pantanowitz, L.: The Ethics of Artificial Intelligence in Pathology and Laboratory Medicine: Principles and Practice. Academic Pathology **8**, 2374289521990784 (Jan 2021). https://doi.org/10.1177/2374289521990784, https://www.sciencedirect.com/science/article/pii/S2374289521000518
9. Kochanny, S.E., Pearson, A.T.: Academics as leaders in the cancer artificial intelligence revolution. Cancer **127**(5), 664–671 (2021). https://doi.org/10.1002/cncr.33284, https://onlinelibrary.wiley.com/doi/abs/10.1002/cncr.33284, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cncr.33284
10. Kretser, A., Murphy, D., Bertuzzi, S., Abraham, T., Allison, D.B., Boor, K.J., Dwyer, J., Grantham, A., Harris, L.J., Hollander, R., Jacobs-Young, C., Rovito, S., Vafiadis, D., Woteki, C., Wyndham, J., Yada, R.: Scientific Integrity Principles and Best Practices: Recommendations from a Scientific Integrity Consortium. Science and Engineering Ethics **25**(2), 327–355 (Apr 2019). https://doi.org/10.1007/s11948-019-00094-3, https://doi.org/10.1007/s11948-019-00094-3
11. Leslie, D.: Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. Tech. rep., Zenodo (Jun 2019). https://doi.org/10.5281/ZENODO.3240529, https://zenodo.org/record/3240529
12. Littmann, M., Selig, K., Cohen-Lavi, L., Frank, Y., Hönigschmid, P., Kataka, E., Mösch, A., Qian, K., Ron, A., Schmid, S., Sorbie, A., Szlak, L., Dagan-Wiener, A., Ben-Tal, N., Niv, M.Y., Razansky, D., Schuller, B.W., Ankerst, D., Hertz, T., Rost, B.: Validity of machine learning in biology and medicine increased through collaborations across fields of expertise. Nature Machine Intelligence **2**(1), 18–24 (Jan 2020). https://doi.org/10.1038/s42256-019-0139-8, https://www.nature.com/articles/s42256-019-0139-8, publisher: Nature Publishing Group
13. Meskó, B., Topol, E.J.: The imperative for regulatory oversight of large language models (or generative AI) in healthcare. npj Digital Medicine **6**(1), 1–6 (Jul 2023). https://doi.org/10.1038/s41746-023-00873-0, https://www.nature.com/articles/s41746-023-00873-0, publisher: Nature Publishing Group
14. Morley, J., Machado, C.C.V., Burr, C., Cowls, J., Joshi, I., Taddeo, M., Floridi, L.: The ethics of AI in health care: A mapping review. Social Science & Medicine **260**, 113172 (Sep 2020). https://doi.org/10.1016/j.socscimed.2020.113172, https://www.sciencedirect.com/science/article/pii/S0277953620303919
15. Morley, J., Morton, C., Karpathakis, K., Taddeo, M., Floridi, L.: Towards a framework for evaluating the safety, acceptability and efficacy of AI systems for health: an initial synthesis (Apr 2021). https://doi.org/10.48550/arXiv.2104.06910, http://arxiv.org/abs/2104.06910, arXiv:2104.06910 [cs]
16. Mousa, A., Flanagan, M., Tay, C.T., Norman, R.J., Costello, M., Li, W., Wang, R., Teede, H., Mol, B.W.: Research Integrity in Guidelines and evIDence synthesis (RIGID): a framework for assessing research integrity in guideline development and evidence synthesis. eClinicalMedicine **74**, 102717 (Aug 2024). https://doi.org/10.1016/j.eclinm.2024.102717, https://www.sciencedirect.com/science/article/pii/S2589537024002967
17. Naik, N., Hameed, B.M.Z., Shetty, D.K., Swain, D., Shah, M., Paul, R., Aggarwal, K., Ibrahim, S., Patil, V., Smriti, K., Shetty, S., Rai, B.P., Chlosta, P., Somani, B.K.: Legal and Ethical Consideration in Artificial Intelligence in Healthcare: Who Takes Responsibility? Frontiers in Surgery **9** (Mar 2022). https://doi.org/10.3389/fsurg.2022.862322, https://www.frontiersin.orghttps://www.frontiersin.org/journals/surgery/articles/10.3389/fsurg.2022.862322/full, publisher: Frontiers

18. Reis-Filho, J.S., Kather, J.N.: Overcoming the challenges to implementation of artificial intelligence in pathology. Journal of the National Cancer Institute **115**(6), 608–612 (Jun 2023). https://doi.org/10.1093/jnci/djad048
19. Seyyed-Kalantari, L., Zhang, H., McDermott, M.B.A., Chen, I.Y., Ghassemi, M.: Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nature Medicine **27**(12), 2176–2182 (Dec 2021). https://doi.org/10.1038/s41591-021-01595-0, https://www.nature.com/articles/s41591-021-01595-0, publisher: Nature Publishing Group
20. Shmatko, A., Ghaffari Laleh, N., Gerstung, M., Kather, J.N.: Artificial intelligence in histopathology: enhancing cancer research and clinical oncology. Nature Cancer **3**(9), 1026–1038 (Sep 2022). https://doi.org/10.1038/s43018-022-00436-4, https://www.nature.com/articles/s43018-022-00436-4, publisher: Nature Publishing Group
21. Siala, H., Wang, Y.: SHIFTing artificial intelligence to be responsible in healthcare: A systematic review. Social Science & Medicine **296**, 114782 (Mar 2022). https://doi.org/10.1016/j.socscimed.2022.114782, https://www.sciencedirect.com/science/article/pii/S0277953622000855
22. Suresh, H., Guttag, J.V.: A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In: Equity and Access in Algorithms, Mechanisms, and Optimization. pp. 1–9 (Oct 2021). https://doi.org/10.1145/3465416.3483305, http://arxiv.org/abs/1901.10002, arXiv:1901.10002 [cs]
23. Vayena, E., Blasimme, A., Cohen, I.G.: Machine learning in medicine: Addressing ethical challenges. PLOS Medicine **15**(11), e1002689 (Jun 2018). https://doi.org/10.1371/journal.pmed.1002689, https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002689, publisher: Public Library of Science
24. Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V.X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M., Ossorio, P.N., Thadaney-Israni, S., Goldenberg, A.: Do no harm: a roadmap for responsible machine learning for health care. Nature Medicine **25**(9), 1337–1340 (Sep 2019). https://doi.org/10.1038/s41591-019-0548-6, https://www.nature.com/articles/s41591-019-0548-6, publisher: Nature Publishing Group

# Assessing the Implementation of the EU 5G Toolbox Recommendations: A Pan-European Study on Cybersecurity Measures in 5G Networks

Diego Santos Alvarez[1], Silvestre Malta[1,2], and Pedro Pinto[3]

[1] Instituto Politécnico de Viana do Castelo, IPVC, Portugal
diegoalvarez@ipvc.pt
[2] ADiT-Lab, Instituto Politécnico de Viana do Castelo, 4900-347 Viana do Castelo, Portugal
smalta@estg.ipvc.pt
[3] GECAD, Instituto Politécnico do Porto, IPP, Portugal
pfp@isep.ipp.pt

**Abstract.** The rapid deployment of 5th Generation (5G) networks across Europe presents significant cybersecurity challenges that must be addressed to ensure the integrity of critical digital infrastructures. The European Union (EU) 5G Toolbox provides strategic and technical recommendations to mitigate risks and strengthen the resilience of 5G networks. This article is based on an ongoing master's thesis project that aims to investigate the implementation of the Toolbox across EU Member States. Although the research is still in progress and does not yet present final results, this paper outlines the objectives and initial scope of the study. The analysis draws on strategic and technical measures described in official documents and academic studies to support the thesis framework and to identify key implementation challenges and future research directions.

**Keywords:** 5G networks · cybersecurity challenges · EU 5G Toolbox · resilience.

## 1 Introduction

The deployment of 5G networks across Europe has introduced unprecedented opportunities for digital innovation, but it has also brought significant cybersecurity challenges. In response, the EU launched the 5G Toolbox—a strategic framework designed to assist member states in enhancing the security and resilience of their national networks. However, the practical implementation of the Toolbox has encountered various obstacles, including economic disparities, dependence on high-risk suppliers, and misalignment between national and EU-level policies.

Despite previous studies addressing specific aspects of 5G infrastructure, there remains a notable research gap regarding the analysis of regional barriers and the identification of best practices to support a harmonized adoption of the Toolbox's recommendations. The rapid expansion of 5G services across the EU has amplified the urgency of addressing these gaps to ensure secure and cohesive network deployment.

This article seeks to fill that gap by analyzing the current state of implementation of the EU 5G Toolbox across Member States. It identifies the key challenges hindering uniform adoption, such as economic and technological asymmetries, and offers strategic and technical recommendations to strengthen Europe's digital infrastructure. These themes also define the core objectives of an ongoing master's thesis, which uses this article as a foundation to explore in depth the barriers to harmonization and the

cybersecurity implications of fragmented implementation. Although final results are not yet available, the article outlines the scope and relevance of the thesis project, serving as an exploratory basis for future academic research and policymaking.

## 2 State of the Art

This section reviews the existing literature on cybersecurity in the context of 5G networks and the implementation of the EU 5G Toolbox. The analysis aims to identify the main approaches and challenges discussed in previous studies, offering an overview of research addressing 5G network security, the policies adopted by the member states of the EU, and the measures proposed to mitigate cybersecurity risks. By presenting relevant works, this section establishes the theoretical foundation and necessary context to understand the evolution and advancements in the implementation of the EU 5G Toolbox.

### 2.1 Key Documents

In [9], the authors present a fundamental document published by the Network and Information Systems (NIS) Cooperation Group in 2020. It describes strategic and technical measures aimed at mitigating risks in 5G networks, emphasizing supplier diversification, strengthening national cybersecurity capabilities, and ensuring resilience through technical standards.

The document [2], released by European Union Agency for Cybersecurity (ENISA) in 2021, complements the Toolbox by providing detailed security requirements and implementation guidelines under the framework of the European Electronic Communications Code (EECC). This document highlights practical steps for member states to adopt measures such as supplier risk assessments and secure network virtualization protocols. The following figure illustrates the overall Toolbox structure along with its corresponding mitigation measures.


**Fig. 1.** 5G Toolbox Structure and Mitigation Measures (source: [2, p. 10]).

The document [4], presented at the ENISA Telecom Security Forum in 2024, focuses on the harmonization of policies among Member States. It underscores the importance of collaborative approaches to mitigate risks and protect critical infrastructures against emerging threats.

### 2.2 Related Work

Several academic studies have addressed 5G cybersecurity from different perspectives. Rogalski emphasizes the importance of defining clear criteria for evaluating supplier security, advocating for comprehensive and rigorous risk assessment processes to strengthen the reliability and resilience of 5G networks [7].

D'Alterio et al. analyze the evolving cybersecurity landscape of 5G networks, offering valuable insights into risk mitigation strategies designed to address emerging and increasingly complex threats [5].

Kadena et al. analyze the technical vulnerabilities inherent in 5G networks in Europe and propose regulatory solutions to close security gaps. Their study highlights the need for rigorous compliance frameworks to ensure the resilience of national infrastructures [3].

Soldani discusses the transformative impact of 5G on security paradigms in ICT, focusing on emerging challenges in the protection and resilience of critical infrastructures. He emphasizes the importance of collaborative approaches among stakeholders to effectively address the cybersecurity threats associated with 5G [1].

Gábriš and Hamuľák explore the relationship between digital sovereignty and cybersecurity, using Slovakia as a case study to illustrate national approaches to 5G security. They argue that digital sovereignty must be understood as the power to control services critical to the state's interests and stress the importance of aligning national strategies with EU legal and technical frameworks such as the 5G Toolbox [13].

Tang et al. conduct a systematic analysis of 5G networks, focusing on core network security. Their study identifies significant architectural challenges and proposes technical solutions to reinforce the protection of 5G core infrastructure, which is essential for the defense of critical systems [10].

Radu and Amon examine risk-based approaches to 5G governance, highlighting the value of effective policy frameworks. They argue that adopting flexible, risk-based strategies can enhance network resilience and support the harmonization of cybersecurity policies across EU member states [11].

Qose and Zoltán also investigate global supply chain challenges in the 5G context, addressing both legal and technical aspects. They emphasize the importance of clear regulation and proactive risk mitigation strategies to reduce dependence on high-risk suppliers [12].

Husić and Baraković provide a practical perspective from telecom operators on the key threats facing 5G networks and countermeasures adopted to mitigate them. The paper discusses actionable security measures to improve the resilience of telecom infrastructure from an operational standpoint [6].

Robles-Carrillo provides a critical examination of EU policy frameworks for establishing secure and resilient 5G infrastructure, identifying significant challenges and shortcomings in their practical implementation across Member States [8].

In summary, the existing literature covers key dimensions of 5G security, including technical vulnerabilities, digital sovereignty, governance models, and supply chain risks. These studies form a foundational basis for understanding the obstacles to harmonized implementation of the EU 5G Toolbox and contribute to the development of future strategies for securing Europe's digital infrastructure.

## 3 Overview of the EU 5G Toolbox

The EU 5G Toolbox, introduced in 2020, comprises a comprehensive set of measures aimed at enhancing the security and resilience of 5G networks [9]. These measures are categorized into two main groups:

### 3.1 Strategic Measures

The implementation of strategic measures is essential to strengthen the security and resilience of 5G networks across the EU. One critical approach is supplier diversification, aimed at reducing dependency on single suppliers, particularly those considered high-risk, to enhance supply chain security. Promoting a balanced market not only fosters competition and innovation but also mitigates potential geopolitical risks.

Another key strategic measure involves rigorous supplier risk assessments. Member states are encouraged to implement thorough assessment protocols to identify both technical and non-technical risks presented by suppliers. These assessments should take into account factors such as the legal obligations that suppliers may have in their country of origin, which could influence their trustworthiness and operational independence.

Additionally, developing national resilience strategies is vital. Member states should create comprehensive contingency plans and ensure the protection of critical network assets. These resilience strategies must be aligned with EU objectives to ensure consistency, interoperability, and overall effectiveness of the cybersecurity framework of the Union.

Together, these strategic measures contribute to a more robust, secure, and resilient European digital infrastructure.

### 3.2 Technical Measures

The adoption of technical measures is fundamental to ensuring the security and resilience of 5G infrastructures across the EU. A primary focus is the definition of minimum security requirements for 5G equipment and networks, encompassing aspects such as software integrity verification and the secure management of Network Function Virtualization (NFV) systems. Establishing these baseline standards is crucial to maintaining consistent levels of protection across all member states.

In addition to security standards, the implementation of standardized certification schemes is proposed. These certification protocols are intended to validate compliance

with EU cybersecurity standards, thus creating a reliable framework to assess and ensure the security of 5G components and systems.

In addition, continuity and resilience planning play a vital role in protecting network operations. Member states are encouraged to develop robust contingency mechanisms and disaster recovery plans to ensure uninterrupted service in the face of disruptions or cyber incidents.

Together, these technical measures aim to establish a harmonized yet adaptable framework capable of protecting 5G infrastructures throughout the EU.

## 4 Challenges in Implementation

This section addresses the main obstacles faced in the implementation of the EU 5G Toolbox by the Member States of the EU. Although the Toolbox provides comprehensive guidelines for strengthening cybersecurity in 5G networks, its application faces a series of economic, political, and technical difficulties that vary significantly between countries. This analysis forms part of a master's thesis currently in progress, and while it does not yet present results, it aims to support the ongoing development of the research.

Among the challenges analyzed, economic disparities are particularly significant. Differences in financial and technical resources between member states hinder the uniform implementation of Toolbox, with smaller economies facing challenges in diversifying suppliers due to cost constraints [3]. In addition to these economic factors, dependence on high-risk suppliers remains a major issue. Some member states are heavily dependent on specific vendors, thus increasing their exposure to supply chain vulnerabilities and geopolitical risks [9].

Coordination challenges also arise when attempting to align national cybersecurity policies with EU strategies. The European Commission report [4] highlights the complexity of reaching consensus on priorities and resource allocation between different stakeholders, illustrating the logistical and political barriers that impede a harmonized approach.

Furthermore, the emergence of sophisticated cyber threats compounds the existing challenges. Recent studies [12], [6] identify new risks such as AI-driven attacks and an increased reliance on international suppliers, emphasizing the urgent need for continuous updates to existing security frameworks.

## 5 Discussion and Future Directions

This section presents a critical analysis of the results and proposes pathways to strengthen the implementation of the EU 5G Toolbox. Based on identified best practices, it highlights effective strategies adopted by Member States and offers recommendations to address ongoing challenges.

Several Member States have successfully implemented the Toolbox by diversifying suppliers and establishing robust regulatory frameworks. Countries such as Germany and France have adopted policies to reduce dependence on high-risk vendors [4], [3],

while the United Kingdom's rigorous security assessments serve as a reference for aligning national practices with EU recommendations [1].

To enhance integration and consistency among Member States, measures such as the creation of centralized platforms for knowledge sharing, joint procurement initiatives to support smaller economies, and targeted financial and technical assistance are suggested. Additionally, continuous adaptation of the Toolbox remains essential to address emerging threats like AI-driven cyberattacks and quantum computing vulnerabilities. These reflections are part of an ongoing master's thesis and are based on publicly available data, with final results still pending.

In this context, further research will focus on developing document-based compliance indicators derived from the analysis of official reports by EU institutions and Member States, enabling comparative assessments of Toolbox implementation. Complementary qualitative and quantitative analyses of national progress reports will help identify gaps, best practices, and recurring patterns. Finally, the construction of a comparative matrix, using information from sources such as the European Commission, ENISA, and the NIS Cooperation Group, will support a broader evaluation of strategic and technical measure adoption across Member States.

It is expected that these approaches will contribute to a more accurate diagnosis of Toolbox implementation and support the development of recommendations to enhance public policies and strengthen the resilience of critical 5G infrastructures within the European Union. Future work will aim to consolidate the collected data, validate the proposed indicators, and advance the comparative analysis.

## 6 Conclusion

This study highlighted the crucial role of the EU 5G Toolbox as a strategic tool to mitigate cybersecurity risks and strengthen the security of 5G networks within the EU. Although it has proven to be a comprehensive and robust framework, its implementation has been marked by significant challenges, particularly due to economic differences between member states and the persistent dependence on high-risk suppliers. This inequality in adoption highlights the need for more coordinated policies and technical and financial support mechanisms that would enable a more uniform application of the proposed guidelines.

The analysis also showed that, in addition to economic disparities, the difficulty of aligning national strategies with EU recommendations hinders a harmonized approach to cybersecurity. Logistical and political challenges still obstruct the building of consensus on security priorities and resource allocation. These difficulties could undermine the resilience of 5G networks across the bloc, especially in light of the growing use of critical technologies that require secure and reliable infrastructures. Therefore, it is recommended to create centralized platforms for sharing best practices and to carry out joint initiatives that could benefit smaller economies.

However, the rapid evolution of cyber threats represents an ongoing risk that necessitates the constant updating of Toolbox's guidelines. With the emergence of more sophisticated attacks driven by artificial intelligence and the looming challenges related to quantum computing, it is imperative to invest in the research and

development of new protection technologies. Future research should focus on promoting greater integration among member states, harmonizing policies and practices, and fostering international partnerships and the adoption of essential technological innovations. In doing so, it will be possible to ensure a secure, resilient digital infrastructure aligned with emerging technological advancements, safeguarding the competitiveness and security of the EU in an increasingly challenging global landscape.

## References


1. David Soldani. “5G and the Future of Security in ICT”. In: 2019 29th International Telecommunication Networks and Applications Conference (ITNAC), Auckland, New Zealand, Nov. 2019, pp. 1-8, doi: 10.1109/ITNAC46935.2019.9078011. url: https://ieeexplore.ieee.org/abstract/document/9078011
2. ENISA. “5G Supplement to the Guideline on Security Measures under the EECC”. In: ENISA Guidelines (July 2021). Detailed guidelines on security measures for 5G infrastructure within the context of the European Electronic Communications Code (EECC). url: https://www.enisa.europa.eu/publications/5g-supplement-security-measures-under-eecc (visited on 11/25/2024).
3. Esmeralda Kadena et al. “5G in Europe: Security and Challenges”. In: 2023 IEEE 17th International Symposium on Applied Computational Intelligence and Informatics (SACI), Timisoara, Romania, May 2023, pp. 000165-000170, doi: 10.1109/SACI58269.2023.10158610. url: https://ieeexplore.ieee.org/abstract/document/10158610
4. European Commission. “EU Policy on the Cybersecurity of 5G Networks”. In: ENISA Telecom and Digital Infrastructure Security Forum 2024. Discussion on cybersecurity policies for 5G networks, including risk mitigation measures for suppliers and protection of critical infrastructures. ENISA, May. 2024. url: https://www.enisa.europa.eu/events (visited on 11/25/2024).
5. Francesco D’Alterio et al. “Navigating 5G Security: Challenges and Progresses on 5G Security Assurance and Risk Assessment”. In: 2024 AEIT International Annual Conference (AEIT), Trento, Italy, Sep. 2024, pp. 1-6, doi: 10.23919/AEIT63317.2024.10736736. url: https://ieeexplore.ieee.org/abstract/document/10736736
6. Jasmina Baraković Husić and Sabina Baraković. “5G Security Threats and Countermeasures: An Operator Perspective”. In: International Conference on Advances in Traffic and Communication Technologies (ATCT). May 2022, pp. 135–140. url:https://atct.ba/wp-content/uploads/2022/06/atct-2022-coference-proceedings.pdf
7. Maciej Rogalski. “Security assessment of suppliers of telecommunications infrastructure for the provision of services in 5G technology”. In: Computer Law and Security Review, Vol. 41, 2021, 105556, ISSN 2212-473X, doi.org/10.1016/j.clsr.2021.105556. url: https://www.sciencedirect.com/science/article/pii/S0267364921000297
8. Margarita Robles-Carrillo. “European Union policy on 5G: Context, scope and limits”. In: Telecommunications Policy, Vol. 45, Issue 8, Sep. 2021, 102216, ISSN 0308-5961, doi.org/10.1016/j.telpol.2021.102216. url: https://www.sciencedirect.com/science/article/pii/S0308596121001208
9. NIS Cooperation Group. “Cybersecurity of 5G Networks: EU Toolbox of Risk Mitigating Measures". Tech. rep. Document presenting the 5G Toolbox, including strategic and technical actions to mitigate cybersecurity risks in the 5G environment, such as supplier diversification and resilience promotion. NIS Cooperation Group, Jan. 2020. url: https://digital-strategy.ec.europa.eu/en/library/cybersecurity-5g-networks-eu-toolbox-risk-mitigating-measures (visited on 11/25/2024).
10. Qiang Tang et al. “A Systematic Analysis of 5G Networks With a Focus on 5G Core Security”. In: IEEE Access, vol. 10, pp. 18298-18319, Feb. 2022, doi: 10.1109/ACCESS.2022.3151000. url: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9709835
11. Roxana Radu and Cedric Amon. “The governance of 5G infrastructure: between path dependency and risk-based approaches”. In: Journal of Cybersecurity, Vol. 7, Issue 1, Aug. 2021, tyab017, doi.org/10.1093/cybsec/tyab017. url: https://academic.oup.com/cybersecurity/article/7/1/tyab017/6350591

12. Silvana Qose and Rajnai Zoltán. “Supply Chain in the Context of 5G Technology Security and Legal Aspects”. In: 2024 IEEE 22nd World Symposium on Applied Machine Intelligence and Informatics (SAMI), Stará Lesná, Slovakia, Jan. 2024, pp. 000143-000148, doi: 10.1109/SAMI60510.2024.10432844.. url: https://ieeexplore.ieee.org/abstract/document/10432844
13. Tomáš Gábriš and Ondrej Hamuľák. “5G and Digital Sovereignty of the EU: The Slovak Way”. In: TalTech Journal of European Studies Tallinn University of Technology (ISSN 2674-4619), Vol. 11, No. 2 (34), doi: 10.2478/bjes-2021-0013. url: https://intapi.sciendo.com/pdf/10.2478/bjes-2021-0013

# Development process of a wearable device for monitoring elderly people with dementia: State of development

Marcelo Arantes[1], Mariana Carvalho[1], Inês C. Rocha[1], Ana Rita Freitas[1], José Soares[1], Marta Pinto[3], Pedro Morais[1,2], Demétrio Matos[4], and Vitor Carvalho[1,2]

[1] 2Ai – School of Technology, IPCA, Barcelos, Portugal
{jarantes, mccarvalho, irocha, arfreitas, jasoares, pmorais, vcarvalho}@ipca.pt
[2] LASI—Associate Laboratory of Intelligent Systems, 4800-058 Guimarães, Portugal
[3] ONECARE, 3030-199 Coimbra, Portugal
mpinto@onecare.pt
[4] ID+, Research Institute in Design, Media and Culture, School of Design, Polytechnic University of Cávado and Ave, 4750-810 Barcelos, Portugal
dmatos@ipca.pt

**Abstract.** The aging global population presents challenges in the management of neurodegenerative diseases such as dementia. This paper describes the development of a wearable system that aims to monitor and support elderly individuals with dementia. Key functional requirements were identified, leading to the creation of an initial functional prototype. The system integrates specific sensors such as BMI270, MAX30101, and SAM-M10Q, as well as the use of artificial intelligence to continuously and accurately monitor health indicators such as heart rate, blood oxygen levels, and daily activities. Focused on comfort and usability, the device is non-intrusive and suitable for continuous wear. Machine learning algorithms process the collected data to enable reliable activity recognition and health monitoring. Although user-based usability tests have not yet been performed, ergonomic and design studies have guided prototype development. The preliminary results of internal technical tests indicate stable sensor data acquisition and successful wireless communication using Bluetooth Low Energy (BLE). These results support the feasibility of the system and prepare the next development phase, which focuses on optimizing the hardware for a compact PCB (Printed Circuit Board) together with the redesign of the three-dimensional model, prioritizing its usability according to the ergonomic studies carried out.

**Keywords:** Wearable Devices · Dementia · Artificial Intelligence · Inclusive Design .

## 1 Introduction

The evolution of wearable technology is increasingly supporting the diagnosis and monitoring of neurodegenerative diseases such as dementia [8]. Dementia causes progressive cognitive decline and early dependence on caregivers, who often face significant physical and emotional challenges [2]. With an aging global population, the prevalence of dementia continues to rise. Therefore, monitoring systems that provide real-time data can help caregivers deliver more responsive and personalized care [10]. The project "Platform for Routine Measurement and Monitoring of Activity in Elderly People with Dementia" aims to develop and validate a new system for monitoring location and daily activities among elderly individuals with dementia, combining technological innovation with a user-centered design approach. This paper is organized into four sections. Section 2 “Methodology”, outlines the approach used

to achieve the stated goals, Section 3 "Development" presents the the progress made in hardware, software, artificial intelligence (AI), and design. Finally, Section 4 "Future Work" describes the next steps planned to advance the system based on the defined methodologies.

## 2 Related work

Recent studies have demonstrated the increasing relevance of wearable technologies in enabling non-invasive, real-time monitoring of vulnerable populations, particularly those at risk for neurodegerative diseases. For example, accelerometer-based devices have been used to assess the impact of moderate physical activity on cognitive performance in individuals with Parkinson's disease, revealing improvements in visuospatial skills, memory, and executive function [6]. Additionally, multi-modal wearable systems have been developed to detect behavioral symptoms of dementia, such as agitation and aggression, with personalized sensor configurations significantly enhancing model performance for activity classification [7]. A recent systematic review titled "Advancing Remote Monitoring for Patients With Alzheimer Disease and Related Dementias" offers a comprehensive evaluation of remote monitoring technologies for dementia care. The review highlights promising developments in wearable devices, home-based sensors, and telehealth platforms, but also emphasizes key challenges such as digital literacy barriers, sensory impairments, and technological accessibility. The authors show the importance of adapting user interfaces and involving caregivers and clinicians early in the design process to ensure adoption, usability, and equity [11].

## 3 Methodology

As part of a wider system represented in Fig. 1 under development, two wearable devices are being developed: and In-ear Wearable Device (HowMI) [5], and a Wrist Wearable Device [9]. This paper focuses on the Wrist Device, which has functional requirements aimed at continuously monitoring heart rate, blood oxygen saturation, step count, sleep patterns, real-time location, falls, and emergency situations. Its development is also guided by four domains of non-functional requirements: physical, emotional, cognitive, and implementation success. Various sensors are used to obtain these monitoring parameters, such as the BMI270 for motion tracking [3], the MAX30101 for heart rate and blood oxygen saturation measurements [1], and the SAM-M10Q for outdoor geolocation via Global Navigation Satellite System (GNSS) [12]. Indoor localization is done through Bluetooth Low Energy (BLE) communication with devices previously placed in known rooms of a building. Data acquisition is managed by embedded software running on a microcontroller, with real-time encrypted data transmission enabled via Bluetooth Low Energy (BLE). The collected data are sent to a nearby base station and then transmitted to a central AI-based processing platform, where machine learning algorithms analyze the data to recognize activity patterns and detect anomalies. Processed information is made available through a secure web portal, allowing caregivers and healthcare professionals

to monitor users remotely and in real time. Fig. 1 illustrates the intended operation of the system, showing the flow of information from the wearable devices to the central AI engine and the interface used by caregivers. The system is designed for deployment in diverse care environments, including senior residences and private homes.

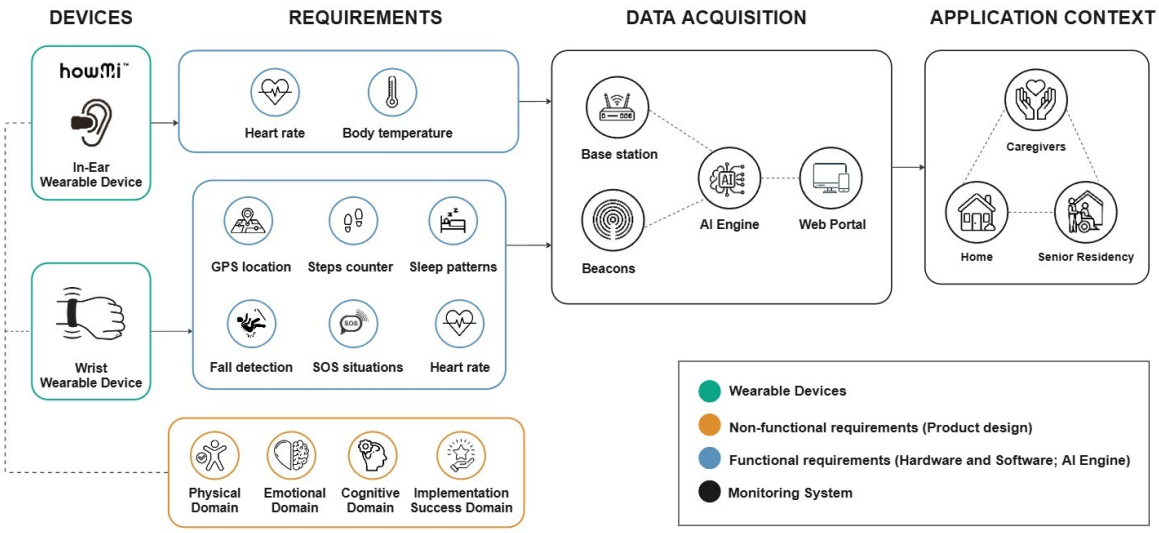



**Fig. 1.** Overview of the project system

## 4 Development

This section outlines ongoing improvements to a wearable device designed for monitoring elderly people with dementia. The development focuses on refining both hardware and software to enhance system performance, user comfort, and sensor integration. Key updates include ergonomic redesigns to improve wearability, expanded data acquisition to support more robust AI model training, upgraded communication protocols for secure and efficient data transmission, and optimized power management to extend device autonomy.

### 4.1 Hardware and Software

A functional prototype of the wearable device was developed using development boards for the selected sensors, organized into three main modules. Each sensor was individually programmed and tested: the BM270 [3] for recording steps, distance, and inertial data; the SAM-M10Q [12] for acquiring geolocation outdoor and time information; and the MAX30101 [1] for high accuracy heart rate and blood oxygen saturation measurement. Indoor location was tested by sending and receiving location-relevant information via BLE protocol. A real-time operating system (RTOS) was used to manage sequential data acquisition, and the BLE protocol was implemented and evaluated for encrypted wireless transmission to external devices.

### 4.2 Usability and Ergonomic Design

Several 3D models and test prototypes were developed to optimize the sensor layout and overall device dimensions, addressing constraints such as mechanical tolerances and space allocation for connectors. After approximately thirty iterations, a functional model was achieved, as shown in Fig. 2. A square or rectangular form factor was initially selected, as it proved more suitable in terms of component arrangement and usability, although special attention was required to ensure comfort during daily use. A smaller screen size was also chosen to match the anatomy of the adult wrist, contributing to a more compact and comfortable device. Final validation will be carried out through usability testing with elderly and their caregivers to gather more accurate and practical feedback.

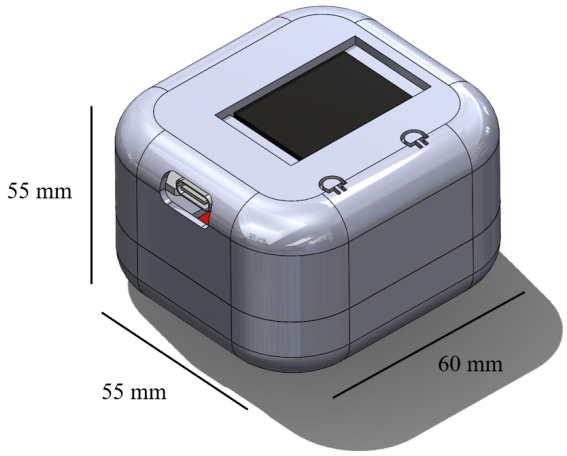



**Fig. 2.** 3D representation of the final functional model with general dimensions

### 4.3 Activities Recognition

AI techniques were developed to support real-time activity recognition and anomaly detection [4]. Machine learning models, such as Support Vector Machines (SVM), Random Forest (RF), K-Nearest Neighbors (KNN), Logistic Regression, Gradient Boosting (XGBoost, LightGBM), were trained on publicly available datasets of wearable sensors to classify activities including walking, sleeping, eating, and hygiene-related tasks. A total of 30 datasets were analyzed, varying in sensor types, activity coverage, participant demographics, and recording duration. Most datasets featured healthy adults, with only a few including elderly individuals or long-term recordings.

Deep learning, including Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM), and Gated Recurrent Units (GRU), were also explored to improve the recognition of more complex behaviors. Training was conducted using an 80/10/10 split for training, validation, and testing and standard metrics including precision, recall,

F1-score, accuracy, and confusion matrices were used for performance evaluation, with macro and weighted-averages reported to account for class imbalance.

Tree-based models, namely XGBoost and LightGBM performed well for common activities and imbalanced datasets, achieving average F1-scores of 0.91 and 0.90, respectively. Deep learning models, particularly GRUs, showed superior accuracy in capturing temporal dependencies and recognizing transitions between complex activities, reaching an F1-score of 0.94 and the highest recall among all models tested.

In addition to activity recognition,this study also explored the application of unsupervised learning approaches to detect deviations from users' typical daily routines. These methods do not rely on labeled anomaly data, making them particularly suitable for healthcare contexts where anomalous events, such as changes in sleep patterns, skipped meals, or irregular mobility, are rare, variable, and difficult to annotate.

Models were trained exclusively on synthetic data representing normal behavior, allowing them to learn baseline routine patterns. Subsequently, deviations were identified based on reconstruction errors or prediction discrepancies, which may reflect early signs of health deterioration, such as increased inactivity, agitation, or disrupted sleep. These AI capabilities enhance the system's ability to monitor daily activities continuously and to provide early alerts to caregivers and healthcare professionals.

Two main strategies were evaluated. The first was a GRU-based autoencoder where anomalies were detected based on reconstruction error, with a threshold defined at the 95th percentile of errors on the training set. A day was flagged as anomalous if more than 10% of its time steps were incorrectly reconstructed. This model achieved an overall accuracy of 92%, a recall of 99% for anomalies, but lower precision (39%), due to a significant number of false positives, especially when the proportion of anomalous intervals per day was low (5%).

The second approach used a GRU-based sequence-to-sequence model with temporal embeddings and an attention mechanism to predict the next activity in a 30-step sequence. Anomalies were identified as mismatches between predicted and actual activity codes, using an error threshold defined as the mean plus two standard deviations of training errors. This method achieved near-perfect precision (99%) in identifying normal routines and high anomaly recall, but also suffered from false positives when anomaly presence was subtle.

While the results are promising, one key limitation remains, the lack of real-world datasets involving individuals with dementia.

## 5 Future Work

The work next step involves analyzing the device's current model to align its technology and design, considering wrist anthropometry to create a final model suitable for the elderly. Key elements like buttons and straps will be studied for usability, with textures and colors differentiating button functions, and a safety system added to the strap to prevent accidental removal. Development will also focus on creating an extensive dataset for monitoring dementia, integrating advanced AI techniques to improve real-time monitoring. Data will be structured and stored via

BLE on smartphones. Efforts will prioritize optimizing data acquisition algorithms, reducing power con-sumption, and enhancing communication protocols, including testing LoRa (Long Range) with AES-CTR (Advanced Encryption Standard - Counter) encryption for security. A hardware redesign will integrate sensors on a compact PCB (Printed Circuit Board). Software improvements will add features, optimize data processing, and enhance energy efficiency, aiming to deliver a cost-effective, user-beneficial product. Although the project is still in development stages, future work will also include usability testing with real users to assess interaction and integration into daily routines. Ethical and regulatory aspects such as informed consent, data privacy, and compliance with relevant standards will also be addressed in collaboration with partner institutions.

## Acknowledgment

This research was funded by the Innovation Pact HfPT—Health From Portugal, co-funded by the "Mobilizing Agendas for Business Innovation" of the "Next Generation EU" program of Component 5 of the Recovery and Resilience Plan (RRP), concerning "Capitalization and Busi ness Innovation", under the Regulation of the Incentive System "Agendas for Business Innovation". This project was also funded through the Foundation for Science and Technology (FCT) under the projects UIDB/05549:2Ai (DOI: 10.54499/UIDB/05549/2020), UIDP/05549:2Ai (DOI: 10.54499/UIDP/05549/2020), CEECINST/00039/2021 and LASI-LA/P/0104/2020.

## References

1. Analog Devices: Max30101 Datasheet / High-Sensitivity Pulse Oximeter and Heart-Rate Sensor for Wearable Health (2020), https://www.analog.com/media/en/technical-documentation/data-sheets/max30101.pdf, analog Devices MAX30101 is an integrated pulse oximetry and heartrate monitor module. It includes internal LEDs, photodetectors, optical elements, and low-noise electronics with ambient light rejection. The MAX30101 provides a complete system solution to ease the design-in process for mobile and wearable devices.
2. Association, A.P.: Diagnostic and Statistical Manual of Mental Disorders (DSM-5-TR), vol. 5 (2022), https://doi/book/10.1176/appi.books.9780890425787
3. Bosh: Bosch BMI270 Datasheet (2023), https://www.bosch-sensortec.com/media/boschsensortec/downloads/datasheets/bst-bmi270-ds000.pdf, data and descriptions in this document are subject to change without notice. Product photos and pictures are for illustration purposes only and may differ from the real product appearance
4. Carvalho, M., Rocha, I., Arantes, M., Linhares, R., Soares, J., Moreira, A., Vilaça, J.L., Matos, D., Morais, P., Carvalho, V.: Powered wearable technologies for dementia care: Evaluating activity recognition models and dataset challenges. In: Proceedings of the 18th International Joint Conference on Biomedical Engineering Systems and Technologies - WHC. pp. 995–1006. INSTICC, SciTePress (2025). https://doi.org/10.5220/0013396600003911
5. Claudino Costa, João M. Faria, D.G.D.M.A.H.M.P.M.J.L.V.V.C.: A wearable monitoring device for covid-19 biometric symptoms detection. Innovation and Research in BioMedical engineering (IRBM) pp. 1–6 (2023), https://doi.org/10.1016/j.irbm.2023.100810
6. Donahue, E.K., Venkadesh, S., Bui, V., Tuazon, A.C., Wang, R.K., Haase, D., Foreman, R.P., Duran, J.J., Petkus, A., Wing, D., Higgins, M., Holschneider, D.P., Bayram, E., Litvan, I., Jakowec, M.W., Van Horn, J.D., Schiehser, D.M., Petzinger, G.M.: Physical activity intensity is associated with cognition and functional connectivity in parkinson's disease. Parkinsonism & Related Disorders **104**, 7–14 (2022). https://doi.org/https://doi.org/10.1016/j.parkreldis.2022.09.005, https://www.sciencedirect.com/science/article/pii/S1353802022002954

7. Iaboni, A., Spasojevic, S., Newman, K., Schindel Martin, L., Wang, A., Ye, B., Mihailidis, A., Khan, S.S.: Wearable multimodal sensors for the detection of behavioral and psychological symptoms of dementia using personalized machine learning models. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring **14**(1), e12305 (2022). https://doi.org/https://doi.org/10.1002/dad2.12305, https://alz-journals.onlinelibrary.wiley.com/doi/abs/10.1002/dad2.12305
8. Robert W. Levenson, Kuan-Hua Chen, D.T.L.E.Y.C.S.L.N.D.P.C.I.Y.C.L.B.J.M.R.D.M., Wang, G.: Evaluating in-home assistive technology for dementia caregivers. Clinical Gerontologist **47**, 78–89 (2024), https://doi.org/10.1080/07317115.2023.2169652
9. Rocha, I., Arantes, M., Moreira, A., Vilaça, J., Morais, P., Matos, D., Carvalho, V.: Monitoring wearable devices for elderly people with dementia: A review. Designs **8**, 75 (07 2024). https://doi.org/10.3390/designs8040075
10. Rodrigues, J.D., Morais, P., Carvalho, V.: Routine measurement and monitoring system for the activity of elderly people with dementia: A systematic review. 9th International Conference on Sensors and Electronic Instrumentation Advances (SEIA' 2023) **20-22 September**, 139–144 (2023),
11. Shaik, M.A., Anik, F.I., Hasan, M.M., Chakravarty, S., Ramos, M.D., Rahman, M.A., Ahamed, S.I., Sakib, N.: Advancing remote monitoring for patients with alzheimer disease and related dementias: Systematic review. JMIR Aging **8**, e69175 (May 2025). https://doi.org/10.2196/69175, https://aging.jmir.org/2025/1/e69175
12. Ublox: SAM-M10Q Datasheet / u-blox M10 standard precision GNSS antenna module (2024), https://content.u-blox.com/sites/default/files/documents/SAM-M10Q_DataSheet_UBX-22013293.pdf, this data sheet describes the SAM-M10Q antenna module with concurrent reception of four GNSS (GPS, Galileo, GLONASS, and BeiDou) and a simple design-in requiring no RF expertise.

# Development of a Modular In-ear Wearable Device for Dementia Monitoring: A Preliminary Approach

Ana Rita Freitas[1], José Soares[1], Inês Rocha[1], Mariana Carvalho[1], Marcelo Arantes[1], Marta Pinto[3], Pedro Morais[1,2], Demétrio Matos[4], and Vítor Carvalho[1,2]

[1] 2Ai – School of Technology, IPCA, Barcelos, Portugal
{arfreitas, jasoares,irocha, mccarvalho, jarantes, pmorais, vcarvalho}@ipca.pt
[2] LASI—Associate Laboratory of Intelligent Systems, 4800-058 Guimarães, Portugal
[3] ONECARE, 3030-199 Coimbra, Portugal
[4] ID+, Research Institute in Design, Media and Culture, School of Design, Polytechnic University of Cávado and Ave, 4750-810 Barcelos, Portugal
dmatos@ipca.pt

**Abstract.** HowMI is a modular in-ear wearable device in early development, designed to enable continuous health monitoring for individuals living with dementia. The device aims to non-invasively measure key vital signs, including heart rate and body temperature, using integrated photoplethysmography and infrared sensors. Design requirements were defined through a systematic literature review and organized into physical, emotional, and cognitive domains, guided by user-centered and inclusive design principles. Based on these requirements, a compact hardware architecture was established, featuring an ESP32 microcontroller and Bluetooth communication for real-time data transmission. A physical prototype is currently under development to test sensor integration and evaluate potential configurations for optimal placement within the ear canal. The HowMI device is conceived as part of a broader modular system, which includes a wrist-worn unit and an AI-based platform for data processing and caregiver support. These initial developments lay the groundwork for future ergonomic studies, sensor performance analysis, and user validation.

**Keywords:** Wearable Device · Dementia · Vital Signs · Inclusive Design

## 1 Introduction

Globally, more than 55 million people live with dementia [7], and an even greater number of caregivers are also affected by the disease. As the condition progresses, the caregiver's role becomes increasingly complex, with physical and emotional exhaustion, isolation, and depression being common [5]. Wearable technology can offer valuable support in monitoring this health condition. This research continues the development of a wearable device called "HowMI - Home Wearables and Monitors Integrated", designed to enhance end-user confidence by providing healthcare professionals with an innovative and more accurate care management tool, enabling faster responses in emergency situations [2]. This short paper presents the technological approach, the preliminary studies, and the future directions of the device in the context of dementia care.

## 2 Methodology

As part of a broader system under development, two types of wearable devices are being created: an In-Ear Wearable Device (HowMI) [2] and a Wrist Wearable Device

[8, 9], each is designed to monitor specific physiological and behavioral parameters. While the overall system integrates both devices, this paper focuses specifically on the preliminary development of the HowMI. To support this system-level approach, the overall architecture is illustrated in Fig. 1.

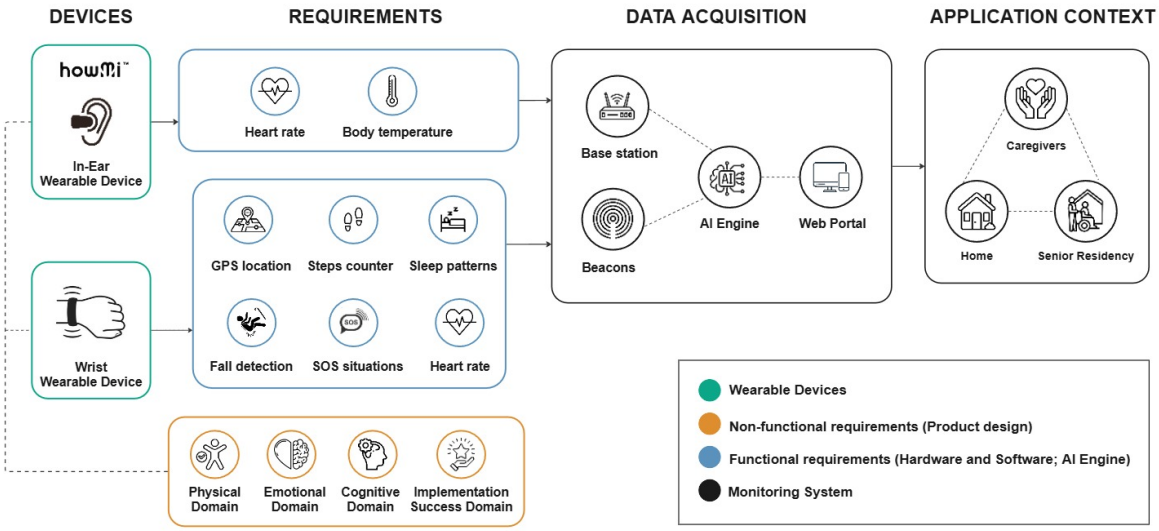



**Fig. 1.** Overview of the project system.

It highlights the integration between the wearable devices and the supporting technological infrastructure. Data collected by the devices is transmitted wirelessly to base stations or beacons, then relayed to a central artificial intelligence (AI) engine [1]. This engine performs real-time data analysis, enabling anomaly detection and generating alerts. Processed information is securely made available through a web portal accessible to caregivers and healthcare professionals. The modular nature of the system ensures that each device can operate independently or in combination, allowing flexible deployment in different care settings, such as private homes or senior residential facilities.

The development of HowMI followed an iterative process guided by user-centered and inclusive design principles, with an emphasis on safety, usability, and efficiency for elderly individuals with dementia. Given the early stage of development, direct involvement of end users was deferred. Instead, initial design requirements were derived from a systematic literature review [3], which synthesized the common needs, preferences, and challenges faced by individuals with dementia and their caregivers. Hardware and software development proceeded in parallel. The current prototype of HowMI integrates temperature and heart rate sensors, selected for their relevance in early detection of physiological changes and their suitability for continuous, non-invasive monitoring in dementia care contexts.

## 3 Preliminary Studies

In this section, we present the preliminary studies that served as the foundation for our project development.

### 3.1 Product Design

To support this research, a systematic literature review was undertaken to explore current trends and challenges in the use of wearable technology for dementia care. The review underscored the emerging potential of ear-based devices such as smart earphones and hearing aids in monitoring key physiological indicators, including heart rate, oxygen saturation, and body temperature. In response to identified limitations in existing solutions, a conceptual design framework was developed, based on user-centered and universal design principles. The resulting requirements were organized into three core domains: physical, emotional, and cognitive. Furthermore, the specific requirements for the HowMi device emphasize the importance of a user-friendly interface, broad adaptability across users, and continuous real-time health data monitoring to enable timely intervention [3].

### 3.2 Hardware

As a first objective in validating the HowMi device, the development and optimisation of the previously designed in-ear device were carried out, with particular emphasis on the hardware. The aim was to reduce the physical dimensions of the device and address any potential limitations without compromising the accuracy of essential biometric data measurements, such as body temperature and oxygen saturation.

As a result of this optimisation, it was possible to reduce the overall length by approximately 1.84 mm, the maximum width by 1.25 mm, and the thickness by 0.6 mm. Two sensors were used, arranged perpendicularly: one infrared sensor, MLX90632 [6], positioned to face the eardrum in order to measure body temperature, and a second photoplethysmography (PPG) sensor, MAX30102 [4], placed in contact with the ear canal to measure heart rate and oxygen saturation.

At this stage, a state-of-the-art review was also conducted to identify the most recent sensory technologies applied in wearable devices, with a particular focus on in-ear solutions. To ensure both relevance and up-to-date information, only articles published from 2019 onwards were selected, written in English and with full free access. The inclusion criteria targeted studies addressing technologies related to the measurement of body temperature, blood oxygen saturation, or heart rate.

This review enabled the exploration of the currently available sensory technologies and a better understanding of their limitations. Two technologies were identified for measuring body temperature in the ear canal: infrared and contact-based. Infrared sensors are limited in that incorrect alignment with the ear canal may lead to inaccurate readings, while contact-based technology requires larger devices that fully occupy the ear canal.

For heart rate and oxygen saturation, photoplethysmography (PPG) technology was observed, which requires contact with the skin for measurement. A notable limitation of this method is that minor movements may interfere with the readings, leading to measurement errors.

The sensors communicate with an ESP32 microcontroller via the $I^2C$ protocol, and the device transmits the data via Bluetooth to a mobile phone.

## 4 Future Work

Future work will focus on advancing the development of the HowMI in-ear wearable device through a series of targeted technical and usability enhancements. A key next step involves conducting a detailed anatomical study of the human ear using 3D models reconstructed from computed tomography (CT) scan data. This analysis will provide critical measurements to optimize the ergonomic fit of the device and ensure user comfort, particularly for long-term use by individuals with dementia. In parallel, a prototype is being developed to support initial testing of the integrated sensors. This prototype will allow the evaluation of sensor performance in an in-ear environment, with particular attention to accuracy, signal stability, and sensitivity to movement. Comparative testing of different sensing technologies and configurations will be conducted to identify the most suitable solutions in terms of precision, energy efficiency, and mechanical compatibility with the ear canal. Further efforts will also address the integration of the in-ear device within the broader HowMI system architecture. This includes refining data transmission protocols, enhancing interoperability with the wrist-worn device, and ensuring seamless communication with the central AI-based data processing platform.

## Acknowledgment

This research was funded by the Innovation Pact HfPT — Health From Portugal, co-funded by the "Mobilizing Agendas for Business Innovation" of the "Next Generation EU" program of Component 5 of the Recovery and Resilience Plan (RRP), concerning "Capitalization and Busi ness Innovation", under the Regulation of the Incentive System "Agendas for Business Innovation". This project was also funded through the Foundation for Science and Technology (FCT) under the projects UIDB/05549:2Ai (DOI: 10.54499/UIDB/05549/2020), UIDP/05549:2Ai (DOI: 10.54499/UIDP/05549/2020), CEECINST/00039/2021 and LASI-LA/P/0104/2020.

## References

1. Carvalho, M., Rocha, I., Arantes, M., Linhares, R., Soares, J., Moreira, A., Vilaça, J., Matos, D., Morais, P., Carvalho, V.: Powered wearable technologies for dementia care: Evaluating activity recognition models and dataset challenges. In: Proceedings of the 18th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 2: WHC. pp. 995–1006. INSTICC, SciTePress (2025). https://doi.org/10.5220/0013396600003911
2. Costa, C., Faria, J.M., Guimarães, D., Matos, D., Moreira, A.H., Morais, P., Vilaça, J.L., Carvalho, V.: A wearable monitoring device for covid-19 biometric symptoms detection. IRBM **44**(6), 100810 (2023). https://doi.org/https://doi.org/10.1016/j.irbm.2023.100810

3. Freitas, A., Soares, J., Arantes, M., Rocha, I., Carvalho, M., Pinto, M., Matos, D., Morais, P., Carvalho, V.: Wearable devices in dementia monitoring: A systematic review of technologies, design, and future directions. EAI Endorsed Transactions on Digital Transformation of Industrial Processes **1**(1) (Apr 2025), https://publications.eai.eu/index.php/dtip/article/view/8856
4. Integrated, M.: Max30102: High-sensitivity pulse oximeter and heart-rate sensor for wearable health, https://www.analog.com/media/en/technical-documentation/data-sheets/MAX30102.pdf
5. Lavretsky, H.: Stress and depression in informal family caregivers of patients with alzheimer's disease. Aging Health **1**, 117–133 (08 2005). https://doi.org/10.2217/1745509X.1.1.117
6. Melexis: Mlx90632 datasheet - infrared sensor, https://www.melexis.com/en/documents/documentation/datasheets/datasheet-mlx90632
7. Organization, W.H.: Global status report on the public health response to dementia: executive summary. World Health Organization (2021)
8. Rocha, I.C., Arantes, M., Moreira, A., Vilaça, J.L., Morais, P., Matos, D., Carvalho, V.: Development of a wearable device for monitoring the activity of elderly people with dementia: first insights. 4th Symposium of Applied Science for Young Researchers: Proceedings (2024)
9. Rocha, I.C., Arantes, M., Moreira, A., Vilaça, J.L., Morais, P., Matos, D., Carvalho, V.: Monitoring wearable devices for elderly people with dementia: A review. Designs **8**(4) (2024). https://doi.org/10.3390/designs8040075, https://www.mdpi.com/2411-9660/8/4/75

# Web-based interactive dashboard with machine learning for Industrial Predictive Maintenance

João Vieira[1], Luis Vilas Boas[1], Inês Caetano[2], Luis Cardoso[3], Paulo Silva[4], Joaquin Dillen[1], João Borges[1], and António H. J. Moreira[1]

[1] 2Ai – School of Technology, IPCA, Barcelos, Portugal
[2] Sistrade, Porto ,Portugal
[3] Miranda, Agueda, Portugal
[4] SRAM, Coimbra, Portugal

**Abstract.** This paper presents an interactive web-based platform for industrial predictive maintenance leveraging machine learning techniques. The system integrates real-time sensor data from manufacturing equipment to provide accurate failure predictions, significantly minimizing unplanned downtime. The technical architecture combines React/Next.js frontend with PostgreSQL database storage and Supabase authentication. Five supervised learning algorithms – Logistic Regression, K-Nearest Neighbors, Support Vector Machines, Random Forest, and XGBoost – were evaluated using the AI4I 2020 Predictive Maintenance dataset, with particular attention to class imbalance addressed through SMOTE resampling techniques. Model performance was comprehensively assessed using metrics as accuracy, Area Under Curve (AUC), F1 score, with hyperparameters tuned via Grid Search to prioritize recall and minimize critical false negatives. The resulting interactive dashboard delivers continuous monitoring of machine health through dynamic charts, visualization tools, and automated alerts. Preliminary results demonstrate this solution's can significantly reduce operational costs, extend equipment lifespan, and improve overall manufacturing efficiency within Industry 4.0 environments.

**Keywords:** Predictive Maintenance · Machine Learning · Industry 4.0 · Web Dashboard · IoT

## 1 Introduction

Predictive maintenance is becoming increasingly important in the industrial sector, especially within the context of Industry 4.0. With the help of digital technologies and continuous equipment monitoring, companies can optimize processes and reduce operational costs. Unlike traditional maintenance strategies that typically fall into two categories: (a) corrective (or reactive) maintenance – often referred to as "fail-and-fix" – where equipment is repaired only after a failure occurs and (b) preventive maintenance, which schedules interventions based on predefined intervals or the expected lifespan of components. While commonly used, these methods can be inefficient, as they may lead to unnecessary part replacements, resulting in wasted resources, or worse, unexpected equipment failures that disrupt production [2], [1].

Predictive Maintenance (PdM) analyzes operational machinery data to determine optimal maintenance windows. Through condition-based monitoring and failure forecasting, PdM enables just-in-time interventions that minimize disruptions while maximizing equipment availability [3].

In this context, the present work introduces a digital platform that predicts industrial machine failures using user-selectable machine learning models and a

web-based dashboard that allows real-time monitoring of sensor data, machine status, and prediction outcomes, enabling proactive maintenance scheduling before breakdowns occur, supporting efficient maintenance planning aligned with Industry 4.0 principles.

## 2 Architecture

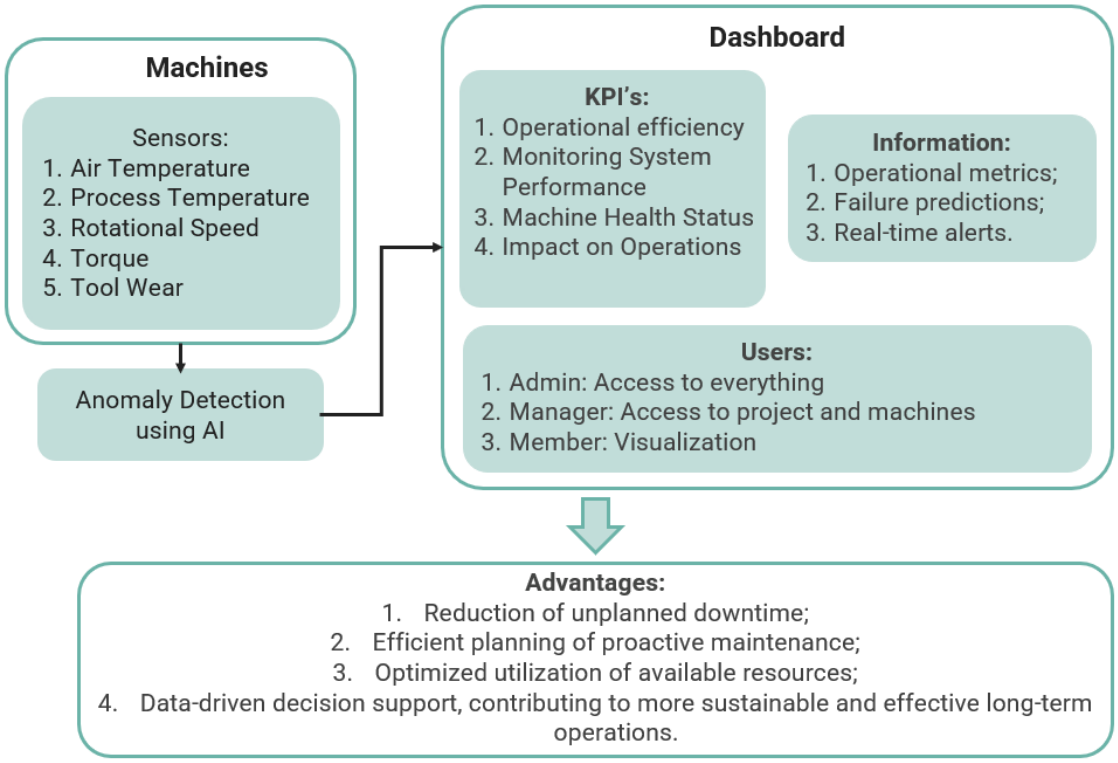



**Fig. 1.** Overview of the predictive maintenance platform, showing sensor data acquisition, AI-based anomaly detection, dashboard features, and operational advantages.

The system, illustrated in Fig. 1, presents a predictive maintenance architecture designed to monitor industrial machines and enable data-driven decisions. Sensor data variables such as air temperature, process temperature, and tool wear are collected in real-time, and then is analyzed using AI-based anomaly detection methods to identify early failure signs. This information feeds into a web-based dashboard displaying real-time KPIs such as operational efficiency and machine health per sector or machine. Developed with modern web technologies, the frontend uses React.js with Next.js App Router and Tailwind CSS, while the backend employs Next.js API Routes. Supabase manages authentication and PostgreSQL database operations via Prisma ORM. Highcharts powers dynamic data visualizations. Model training and validation and testing used the AI4I 2020 dataset from the UCI Repository with a split of 80/10/10, respectively. The ML pipeline includes data analysis and

preprocessing with Pandas, NumPy, Matplotlib, and Seaborn, while Scikit-Learn and Imbalanced-Learn (with SMOTE) handled modeling and class imbalance. Evaluated models included Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest, and XGBoost.

## 3 Results

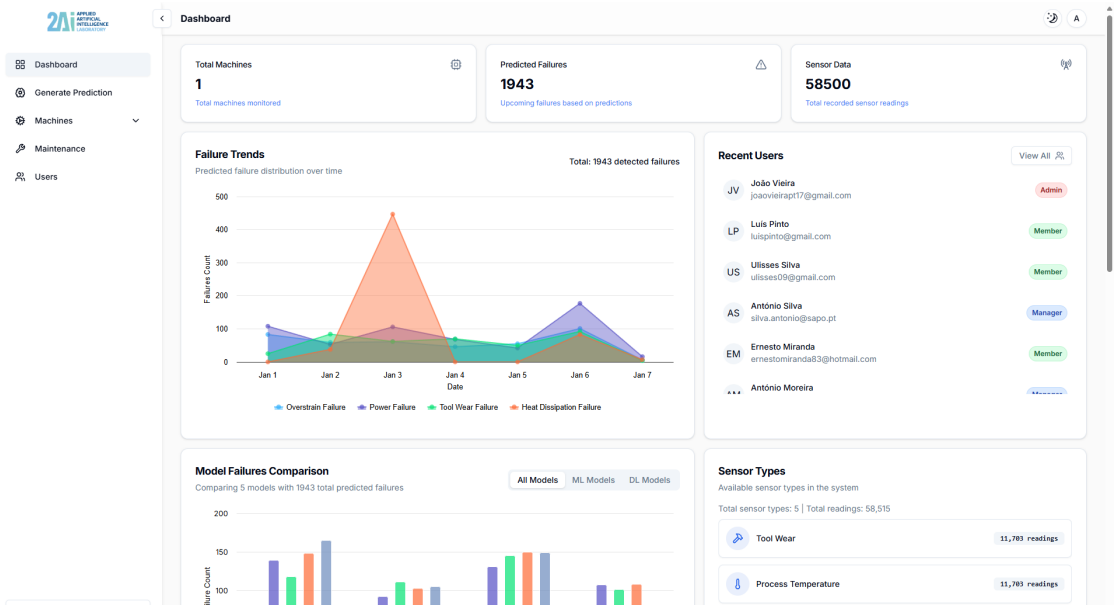



**Fig. 2.** Example of the dashboard presented to the user, showing model predictions, machines health, and risk of eminent failure.

The implemented predictive maintenance dashboard, shown in Fig. 2, offers an intuitive interface for monitoring machine health and predicting potential failures. The system tracks multiple machines and forecasts failures across all units. Four charts like Failure Trends, most common failure, mahcine Health and Model Failure Comparison display key information, such as failure trends and historical data, allowing users to monitor machine conditions over time. These trends help detect early signs of potential failures if certain metrics deviate from normal patterns, enabling proactive interventions.

The dashboard also includes a section that visualizes the distribution of common failure types, providing an overview of the most frequent failure points. This allows maintenance teams to anticipate and prioritize issues effectively.

The machine health status section delivers real-time risk assessments, accompanied by actionable recommendations. This feature supports proactive maintenance planning by forecasting failures and suggesting maintenance actions up to several days in advance.

As shown in Table 1, the results demonstrate strong performance across all models, with XGBoost consistently achieving the highest scores (ACC: 0.982-0.985, AUC:

**Table 1.** Model Performance Comparison

| | Metrics | KNN | SVC | RFC | XGB |
|---|---|---|---|---|---|
| **Validation** | ACC | 0.956 | 0.968 | 0.978 | 0.982 |
| | AUC | 0.956 | 0.993 | 0.998 | 0.999 |
| | F1 | 0.957 | 0.969 | 0.978 | 0.982 |
| | F2 | 0.957 | 0.968 | 0.978 | 0.982 |
| **Parameters** | Neighboors | 1 | - | - | - |
| | C | - | 100 | - | - |
| | Gamma | - | 1 | - | - |
| | Learning Rate | - | - | - | 0.1 |
| | Nº Estimators | - | - | 700 | 500 |
| | Max Depth | - | - | 10 | 5 |

0.999). Random Forest closely follows as the second-best performer. While KNN and Logistic Regression show lower metrics, they still maintain respectable accuracy above 0.92. All models demonstrate good generalization with minimal differences between validation and test scores. XGBoost emerges as the optimal choice for this classification task, offering the best performance across all evaluation metrics.

## Acknowledgement

This research was carried out in partnership with the companies Sistrade - Software Consulting, SA, SRAM and Miranda & Irmão, Lda, funded by the "Agenda Mobilizadora para a inovação empresarial do setor das Duas Rodas (AM2R)" 01/C05-i01/2021 and LASI-LA/P/0104/2020, co-funded from the "Mobilising Agendas/Alliances for Business Innovation" of the "Next Generation EU" program of Component 5 of the Portuguese Recovery and Resilience Plan (RRP), concerning "Investment and innovation", under the Regulation of the Incentive System for "Mobilising Agendas/Alliances for Business Innovation".

## References

1. Kamat, P., Sugandhi, R.: Anomaly detection for predictive maintenance in industry 4.0-a survey. In: E3S web of conferences. vol. 170, p. 02007. EDP Sciences (2020)
2. Li, X., Ding, Q., Sun, J.Q.: Remaining useful life estimation in prognostics using deep convolution neural networks. Reliability Engineering & System Safety **172**, 1–11 (2018). https://doi.org/https://doi.org/10.1016/j.ress.2017.11.021, https://www.sciencedirect.com/science/article/pii/S0951832017307779
3. Ruiz-Sarmiento, J.R., Monroy, J., Moreno, F.A., Galindo, C., Bonelo, J.M., Gonzalez-Jimenez, J.: A predictive model for the maintenance of industrial machinery in the context of industry 4.0. Engineering Applications of Artificial Intelligence **87**, 103289 (2020)

# Reinforcement Learning-Driven Autonomous Cyber Defense Agents for Adaptive Incident Response

Safa Ouerchfani[1,2], Rui Fernandes[1,3,4], and Nuno Lopes[1,3]

[1] 2AI - School of Technology - IPCA, Barcelos, Portugal
a51039@alunos.ipca.pt, rpfernandes@ipca.pt, nlopes@ipca.pt
[2] ISI - Higher Institute of Computer Science, Tunis, Tunisia
[3] LASI – Associate Laboratory of Intelligent Systems, Portugal
[4] TUS - Technological University of the Shannon, Limerick, Ireland

**Abstract.** The increasing complexity and frequency of cyber threats require innovative and adaptive approaches to Incident Response. Traditional methods often fail to address the speed and sophistication of modern cyberattacks. Intelligent and autonomous defence mechanisms are critical as adversaries employ more dynamic and evasive techniques. This paper explores the Autonomous Cyber Defense Agents concept, focusing on those that utilise Reinforcement Learning (RL) to facilitate Dynamic Incident Response. RL enables agents to learn optimal strategies through continuous interaction with their environments, allowing them to adapt quickly to emerging threats. By designing and using realistic, simulated cyber environments for training, organisations can equip RL agents with the capabilities required for proactive threat detection, decision-making, and response execution. The proposed approach aims to bridge the gap between current reactive security practices and the future of self-learning autonomous cyber defence.

**Keywords:** Incident Response · Reinforcement learning · Autonomous Defense Agents · Cybersecurity

## 1 Introduction

The increasing sophistication and frequency of cyberattacks have highlighted the need for more adaptive and intelligent approaches to Incident Response. Traditional methods, which rely on predefined procedures and manual interventions, are increasingly inadequate against modern threats such as Advanced Persistent Threats (APTs) and zero-day exploits [4]. These evolving challenges require faster, more flexible, and automated response mechanisms.

Artificial Intelligence (AI) is increasingly being leveraged in Incident Response to improve threat detection, accelerate reaction times, and automate routine tasks [6]. Machine learning techniques assist in identifying anomalies, classifying threats, and triaging alerts, while AI-driven analysis supports root cause identification and post-incident learning. Despite these advancements, many AI methods rely on static models or predefined labels, limiting their ability to adapt to rapidly evolving cyber threats [1].To address these limitations, tools like Cyborg integrate AI with defensive capabilities such as automated detection, adaptive threat response, and behavioral analysis, offering a more dynamic and resilient approach to Incident Response [5].

However, Cyborg presents several notable limitations. It does not support the creation of diverse network topologies at scale and lacks sufficient red agent diversity and specificity, which reduces the realism of attack scenarios. Additionally, it lacks

visualization tools necessary to observe agent behavior—an essential feature for explainability and for assessing and diagnosing potential issues in agent performance [12]. In contrast, CyberWheel offers a more complex and flexible environment that addresses these shortcomings, enabling the training of more robust and adaptable defensive agents.

Recent advances in Artificial Intelligence, particularly Reinforcement Learning (RL), offer promising new directions to enhance cyber defence capabilities [13].
It addresses the gap by enabling agents to learn optimal response strategies through continuous interaction with dynamic environments. Unlike traditional models, RL allows systems to autonomously adapt to new attack patterns, making it a promising approach for real-time, intelligent Incident Response.

The paper is organised as follows: Section 2 provides background on Incident Response and Reinforcement Learning. Section 3 introduces the structure of the CyberWheel environment. Section 4 outlines our future work.

## 2 Background

The traditional Incident Response (IR) cycle, according to the National Institute of Standards and Technology (NIST), follows a linear, predefined process consisting of three major phases as shown in the Fig. 1: Detect, Respond, and Recover [10].


**Fig. 1.** Cyber Incident Response Cycle

While this structured approach has long provided a foundation for managing cyber incidents, it increasingly shows its limitations in today's rapidly evolving threat landscape. This cycle heavily depends on human analysts and predefined playbooks, which often lack the agility needed to respond to novel, fast-moving threats [15]. As threats become more dynamic and adaptive, the rigidity of this cycle, combined with the reliance on tools like signature-based intrusion detection systems (e.g., Snort, Suricata), renders many IR processes too slow or outdated to be effective in real time [14]. As highlighted in [8], such static, rule-based systems fail to detect zero-day

attacks or advanced multi-stage intrusions, thereby exposing organisations to prolonged dwell times and increased damage before threats are fully neutralised.

These challenges highlight the need for a more adaptive and intelligent Cybersecurity approach. As threats grow more complex and digital environments become more dynamic, traditional methods fall short. This motivates the use of Reinforcement Learning to develop autonomous agents capable of learning from and responding to threats in real time, offering a scalable and proactive alternative to manual Incident Response [14].

Reinforcement learning (RL), a branch of Machine Learning, offers a powerful framework for building intelligent agents capable of making sequential decisions in dynamic and uncertain environments [9]. Unlike traditional supervised learning methods, RL does not rely on labelled datasets; instead, it allows agents to learn through interaction with their environment by taking actions and receiving feedback in the form of rewards. This trial-and-error process enables the agent to gradually learn an optimal policy – a mapping from observed states to actions – that maximises long-term rewards [9].

In Cybersecurity contexts, RL presents unique advantages. As attack surfaces expand and threats evolve, defenders often operate with limited prior knowledge about the environment or attacker behaviour [11]. RL addresses this challenge by allowing autonomous agents to adaptively explore and exploit their surroundings, enabling real-time decision-making in unfamiliar scenarios. RL agents can also operate in high-dimensional state spaces and manage both discrete and continuous actions, making them suitable for complex tasks such as threat detection, intrusion prevention, and response automation [16].

The feedback mechanism central to RL, where agents receive positive or negative reinforcement based on their actions, supports continuous adaptation. In security applications, this enables agents to improve performance over time by learning which actions effectively counter threats, even when the optimal response is not explicitly known. This continuous learning process enhances the agent's ability to cope with evolving attack techniques [3].

## 3 RL-based agent for Incident Response

CyberWheel [12] is a training environment that provides simulation capabilities. It enables flexible customisation of training scenarios, allowing users to easily redefine key components such as the agent's reward function, observation space, and action space. This adaptability supports rapid experimentation with new agent designs. Additionally, the platform offers valuable insights into agent behaviour for effective evaluation and includes comprehensive documentation and examples to facilitate adoption by researchers and practitioners [12].

Cyberwheel's simulation environment is built on a robust network definition code and configuration files that enable rapid experimentation across topology sizes and configurations. Its network is comprised of routers, subnets, and hosts, each of which is represented, in Fig. 2 below, as a node on a networkx graph [12]. The figure describes a network topology of 3 subnets: hosts, servers and DMZ. It indicates that the node

colours correspond to the red agent's actions: yellow represents reconnaissance, green indicates discovery, and grey signifies the absence of any action [12].

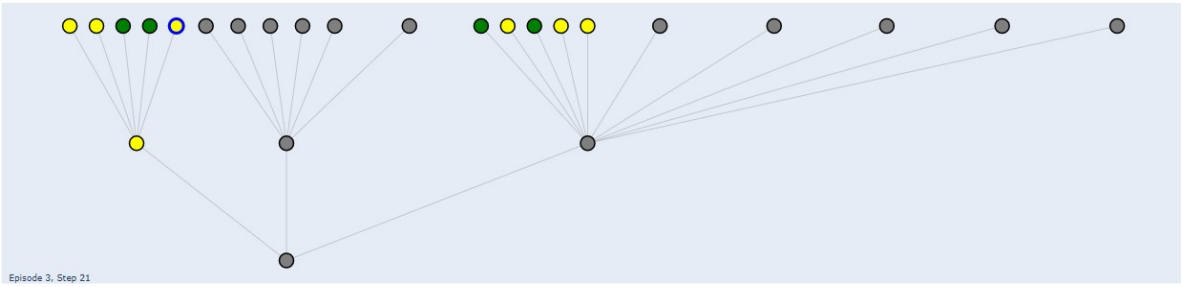



**Fig. 2.** Cyberwheel Network Architecture

CyberWheel is relatively recent and its full capabilities have not yet been extensively evaluated. To address this gap, we propose a systematic evaluation of CyberWheel by testing all its available defensive actions across various simulated environments and under different Reinforcement Learning algorithms, CyberWheel is relatively recent, and its full capabilities have not yet been extensively evaluated. To address this gap, we propose a systematic evaluation of CyberWheel by testing all its available defensive actions across various simulated environments and under different Reinforcement Learning algorithms. Specifically, we compare the performance of three widely used RL algorithms – Proximal Policy Optimization (PPO) [7], Q-Learning [2], and Deep Q-Networks (DQN) [7] – to assess their effectiveness in training CyberWheel agents. This approach aims to assess the simulator's flexibility, effectiveness, and suitability for training adaptive Incident Response agents in realistic threat scenarios, leveraging Reinforcement Learning's ability to learn optimal defense strategies through trial and error based on feedback from reward signals.

## 4 Future work

As future work, we plan to benchmark the CyberWheel simulator to assess its feasibility as a reliable environment for training and evaluating Reinforcement Learning-based Cybersecurity agents. This includes a comparative analysis of three RL algorithms—PPO, Q-Learning, and DQN—focusing on key performance indicators such as training time and cumulative rewards obtained across different scenarios. The goal is to determine CyberWheel's practicality and effectiveness for broader research use in Autonomous Incident Response,as well as to identify which type of Reinforcement Learning algorithm is best suited for this specific use case.

## References

1. Ali, S., Wang, J., Leung, V.C.M.: Ai-driven fusion with cybersecurity: Exploring current trends, advanced techniques, future directions, and policy implications for evolving paradigms–a comprehensive review. Information Fusion p. 102922 (2025)

2. Chadi, M.A., Mousannif, H.: Understanding reinforcement learning algorithms: The progress from basic q-learning to proximal policy optimization. arXiv preprint arXiv:2304.00026 (2023)
3. Dunsin, D., Ghanem, M.C., Ouazzane, K., Vassilev, V.: Reinforcement learning for an efficient and effective malware investigation during cyber incident response. High-Confidence Computing p. 100299 (2025)
4. Dupont, B., Shearing, C., Bernier, M., Leukfeldt, R.: The tensions of cyber-resilience: From sensemaking to practice. Computers & security **132**, 103372 (2023)
5. Emerson, H., Bates, L., Hicks, C., Mavroudis, V.: Cyborg++: An enhanced gym for the development of autonomous cyber agents. arXiv preprint arXiv:2410.16324 (2024)
6. Fernandes, R., Lopes, N., Gonçalves, J.: Integration of ai and embedded systems for enhanced intrusion detection system: A case study with nvidia jetson nano. In: 4th Symposium of Applied Science for. p. 29 (2024)
7. Neil de la Fuente, D.A.V.G.: A comparative study of deep reinforcement learning models: Dqn vs ppo vs a2c. arXiv preprint arXiv:2407.14151 (2024)
8. Kasowaki, L., Kaan, M.: The evolving threatscape: understanding and navigating cybersecurity risks. EasyChair Preprint #11703 (2024), `https://easychair.org/publications/preprint/MxXq/open`
9. Madsen, H., Grov, G., Mancini, F., Baksaas, M., Sommervoll, Ć., et al.: Exploring reinforcement learning for incident response in autonomous military vehicles. arXiv preprint arXiv:2410.21407 (2024)
10. NIST Incident Response Project Team: Incident response,overview (2025), `https://csrc.nist.gov/projects/incident-response`
11. Oesch, S., Austria, P., Chaulagain, A., Weber, B., Watson, C., Dixson, M., Sadovnik, A.: The path to autonomous cyberdefense. IEEE Security & Privacy (2024)
12. Oesch, S., Chaulagain, A., Weber, B., Dixson, M., Sadovnik, A., Roberson, B., Watson, C., Austria, P.: Towards a high fidelity training environment for autonomous cyber defense agents. In: Proceedings of the 17th Cyber Security Experimentation and Test Workshop. pp. 91–99 (2024)
13. Palmer, G., Parry, C., Harrold, D.J., Willis, C.: Deep reinforcement learning for autonomous cyber defence: A survey. arXiv preprint arXiv:2310.07745 (2023)
14. Ren, S., Jin, J., Niu, G., Liu, Y.: Arcs: Adaptive reinforcement learning framework for automated cybersecurity incident response strategy optimization. Applied Sciences **15**(2), 951 (2025)
15. Schlette, D., Empl, P., Caselli, M., Schreck, T., Pernul, G.: Do you play it by the books? a study on incident response playbooks and influencing factors. In: 2024 IEEE Symposium on Security and Privacy (SP). pp. 3625–3643. IEEE (2024)
16. Tareq, I., Elbagoury, B.M., El-Regaily, S.A., El-Horbaty, E.S.M.: Deep reinforcement learning approach for cyberattack detection. International Journal of Online & Biomedical Engineering **20**(5) (2024)

# LLM-PentestBot: A RAG-Based Assistant for Penetration testing tasks

Oumaima Ben Fadhel[1,2], Rui Fernandes[1,3,4], and Nuno Lopes[1,3]

[1] 2Ai – School of Technology, IPCA, Barcelos, Portugal
a51039@alunos.ipca.pt, {rpfernandes, nlopes}@ipca.pt
[2] ISI - Higher Institute of Computer Science, Tunisia
[3] LASI – Associate Laboratory of Intelligent Systems, Portugal
[4] TUS - Technological University of the Shannon, Limerick, Ireland

**Abstract.** Penetration testing, also referred to as Ethical Hacking, is a fundamental practice in Cybersecurity that involves evaluating the security of systems, networks, and applications to identify vulnerabilities and potential attacks. Traditional approaches to Penetration testing are often time-consuming, rely on manual processes, and require advanced technical expertise. This paper presents LLM-PentestBot, a Chatbot assistant that uses Large Language Models (LLMs) integrated with a Retrieval-Augmented Generation (RAG) pipeline to support and automate various Penetration testing tasks. The assistant facilitates tasks such as vulnerability enumeration, command generation by dynamically retrieving relevant domain-specific knowledge from different sources, and results interpretation. This approach enhances the efficiency and effectiveness of Penetration testing for both professionals and beginners.

**Keywords:** Penetration testing · Large-Language-Models · Retrieval-Augmented-generation · Cybersecurity

## 1 Introduction

With the increasing adoption of digital technologies, the explosion in data touchpoints has increased vulnerabilities. The National Vulnerability Database (NVD) recorded over 30,000 new Common Vulnerabilities and Exposures (CVEs), half classified as high or critical severity. In connection with this rise in vulnerabilities, the NVD has also reported that the global average cost of a data breach in 2024 reached $4.88 million, a 10% increase from the previous year [7]. Artificial Intelligence (AI) has been adopted across various domains and is now incorporated into Cybersecurity practices such as Penetration testing [8]. Recent works like PentestGPT [3] and AutoPentest [5] explore the use of LLMs in Penetration testing. PentestGPT, based on GPT-4, assists in CTF and web scenarios using modular agents but lacks retrieval capabilities, making it prone to hallucinations [2] and context loss over long task sequences. On the other hand, AutoPentest integrates GPT-4o within a LangChain-based multi-agent framework featuring RAG capabilities which enhances context awareness and accuracy during testing. However, it only partially automates tasks (completing about 15–25% of tests autonomously) and relies on costly API usage ( $96 per run), limiting scalability and requiring human oversight for complex decisions.

In this context, we propose the development of LLM-PentestBot, an open-source AI-driven assistant with no API costs that combines LLMs with an RAG pipeline to support Penetration testing tasks.

This paper is structured as follows: Section 1 is an introduction to the study of the existence of Penetration testing and artificial intelligence in today's world. Section 2 provides the necessary background information to support the understanding of the topic. Section 3 presents an overview of the system, explaining its architecture and outlining the expected outcomes. Finally, Section 4 discusses directions for future work.

## 2 Background

Penetration tests are security tests that imitate real attacks to identify methods and ways of bypassing the security measures of a system or a network. This often involves running real attacks on real systems and data using specialized tools and techniques commonly used by attackers. It follows a structured process composed of several key phases [10] as presented in Fig. 1.

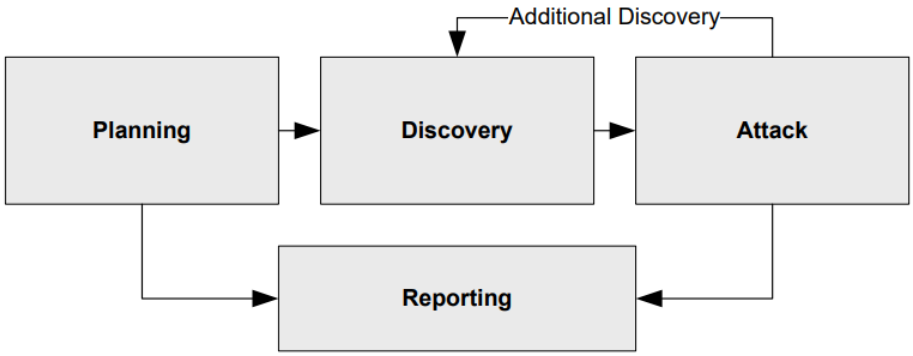



**Fig. 1.** Penetration testing methodology diagram

**Planning:** In this phase, the test scope is defined, and necessary approvals and legal agreements are obtained. The team gathers key information about the organization's operations and security policies to prepare for the test.

**Discovery:** Also called information gathering, this phase involves identifying systems, services, and potential vulnerabilities using both non-intrusive and intrusive methods.

**Exploitation:** Using the data from discovery, the team attempts to exploit vulnerabilities to assess the impact of real-world attacks. This phase may loop back to discovery as new information emerges.

**Reporting:** Throughout the test, findings are documented. At the end, a detailed report is delivered, outlining vulnerabilities, risk levels, and recommendations for remediation.

The attack phase in Penetration testing involves four main steps: gaining access to the target system, escalating privileges to obtain deeper control, browsing the system to explore resources, and installing additional tools to maintain access or extract data.

Effective Penetration testing traditionally requires the use of various specialized tools, each designed for specific tasks such as network scanning, vulnerability assessment, and exploitation. This tool-based approach requires significant technical expertise and manual effort, creating a high barrier to entry, making traditional pentesting resource-intensive and highly dependent on the expertise of the tester.
To address these challenges, Artificial Intelligence and particularly Large Language Models (LLMs) is increasingly being integrated into the Penetration testing process. LLMs, a form of generative AI, are foundation models trained on vast amounts of data, enabling them to understand and generate natural language and other content to support a wide range of tasks [6].
Retrieval-Augmented Generation (RAG) is a recent approach designed optimize LLM outputs by incorporating information from an external, authoritative knowledge base at query time, rather than relying solely on the model's pre-trained data. [1].

The combination of LLMs with RAG allows us to build AI systems capable of generating and interpreting human-like text. By setting a knowledge base of Penetration testing-related content, we can aim for the development of a Penetration testing assistant, increasing its performance.

## 3 System Proposal

In this project, we propose the development of a Chatbot assistant designed to help penetration testers throughout the entire testing process. The assistant uses LLMs enhanced with an RAG pipeline [1] to provide contextual real-time guidance. By combining dynamic access to a Cybersecurity knowledge database with natural language interaction, this assistant aims to simplify complex tasks, reduce the learning curve for newcomers, and improve efficiency and consistency in ethical hacking operations.

### 3.1 System Architecture

RAG is an advanced architectural framework that enhances the performance of LLMs by integrating them with external dynamic knowledge sources. Unlike traditional LLMs limited by static training data, RAG enables real-time retrieval of relevant information during inference, thus improving the contextual relevance of responses. This design significantly reduces the risk of model hallucinations and supports more reliable outputs. Furthermore, by separating the knowledge base from the model itself, RAG allows continuous updates without the need for retraining, ensuring scalability and long-term adaptability. Fig. 2 illustrates the general architecture of the proposed LLM-based chatbot assistant.

The RAG pipeline offers strong flexibility by integrating diverse data types from various sources. Our system is designed to extract textual content from PDF files while also handling images embedded within them. When images such as command-line screenshots contain relevant textual information, our system applies Optical Character Recognition (OCR) techniques [9] to extract text, use the LLM model to only keep the relevant extracted content and integrate it into the retrieval

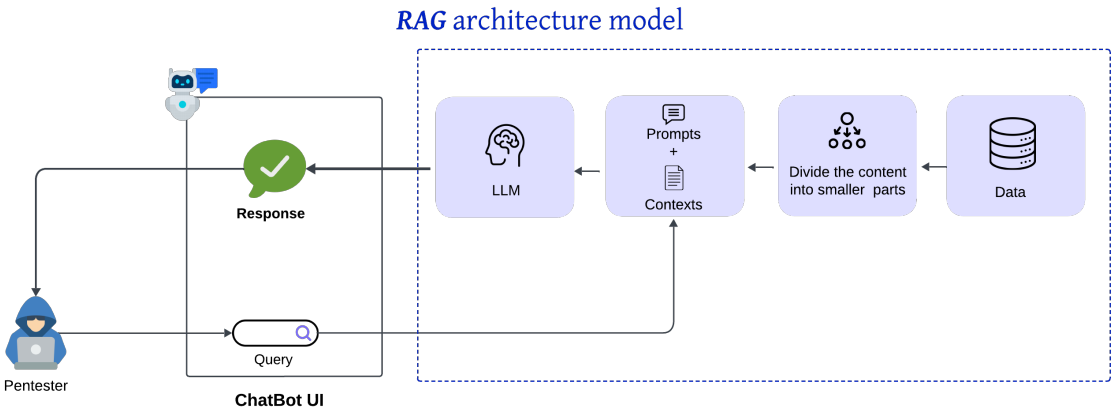



**Fig. 2.** Chatbot architecture

index. This multimodal input enables the system to provide accurate, context-aware guidance based on real-time, comprehensive knowledge.

Moreover, our system applies a layered prompt engineering strategy [4] to define instruction for the LLM to behave as an ethical hacking assistant and to provide clear and structured answers to user queries, ranging from general questions and vulnerability analysis to summarized report generation. Some of the defined prompts were fed with examples or pairs of input/output to guide the LLM for the desired output format.

### 3.2 Expected Results

The proposed Chatbot delivers a range of outputs to support pentest activities and facilitate comprehensive evaluation, as outlined in Table 1.

**Table 1.** Expected Outputs of the LLM-Based Chatbot Assistant

| Output Type | Description |
|---|---|
| General Cybersecurity Answers | Responses to common questions in the Cybersecurity and ethical hacking domains. |
| Command Suggestions | Provides relevant CLI commands for each phase of the penetration test. |
| Output Analysis | Interprets and explains the results from various pentesting tools. |
| Mitigation Recommendations | Suggests actions to remediate identified vulnerabilities. |
| Report Generation | Produces structured summaries of findings, including exploited issues and fixes. |

## 4 Future Work

As future work, we plan to implement the proposed system based on an RAG architecture, incorporating textual and visual data sources. Our objective is to develop a Chatbot capable of providing accurate and up-to-date responses to

pentest-related queries. Additionally, we intend to compare the performance of different LLM models to identify the most effective configuration for the system.

## References


1. AWS: What is rag (retrieval-augmented generation)? (2025), https://aws.amazon.com/what-is/retrieval-augmented-generation/
2. Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q.V., Xu, Y., Fung, P.: A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. Centre for Artificial Intelligence Research (CAiRE) The Hong Kong University of Science and Technology **80**(45), 2 (2023). https://doi.org/https://arxiv.org/pdf/2302.04023
3. Deng, G., Liu, Y., Mayoral-Vilches, V., Liu, P., Li, Y., Xu, Y., Zhang, T., Liu, Y., Pinzger, M., Rass, S.: PentestGPT: Evaluating and harnessing large language models for automated penetration testing. In: 33rd USENIX Security Symposium (USENIX Security 24). pp. 847–864. USENIX Association, Philadelphia, PA (Aug 2024), https://www.usenix.org/conference/usenixsecurity24/presentation/deng
4. Ekin, S.: Prompt engineering for chatgpt: A quick guide to techniques, tips, and best practices (05 2023). https://doi.org/10.36227/techrxiv.22683919
5. Henke, J.: Autopentest: Enhancing vulnerability management with autonomous llm agents (2025), https://arxiv.org/abs/2505.10321
6. IBM: What are large language models (llms)? 5 (2023), https://www.ibm.com/think/topics/large-language-models
7. (NVD), N.V.D.: Key cyber security statistics for 2025 (2025), https://www.sentinelone.com/cybersecurity-101/cybersecurity/cyber-security-statistics/
8. Pecan: The role of llms in ai innovation, https://www.pecan.ai/blog/role-of-llm-ai-innovation/
9. Saoji, S., Eqbal, A., Vidyapeeth, B.: Text recognition and detection from images using pytesseract. J Interdiscip Cycle Res **13**, 1674–1679 (2021)
10. Scarfone, K., Souppaya, M., Cody, A., Orebaugh, A.: Technical guide to information security testing and assessment. NIST Special Publication **800**(115), 2–25 (2008)

# Thermal Monitoring of Aluminum Forging for Process Optimization

Joaquin Dillen[1], Luis Vilas Boas[1], N. Simões[1], João M. Faria[1], Rafael Fernandes[1], Bruno Silva[1], Inês Caetano[2], João Borges[1], and António H. J. Moreira[1]

[1] 2Ai – School of Technology, IPCA, Barcelos, Portugal
{jdillen, lvilasboas, nsimoes, jfaria, rmfernandes, brsilva, jpbsilva, amoreira}@ipca.pt
[2] Sistrade, Software Consulting, SA, Porto, Portugal
ines.caetano@sistrade.pt

**Abstract.** Achieving consistent product quality in forged aluminum components remains a challenge due to thermal variability introduced during the quenching phase, which can lead to residual stress and surface defects. This paper presents a thermal imaging-based monitoring system designed to capture real-time surface temperature data during cooling, enabling early detection of thermal inconsistencies. A prototype system was implemented and evaluated in a production environment, successfully capturing over 4,000 thermal images across four consecutive manufacturing days. Analysis revealed that the quenching fluid maintained a stable thermal profile with temperature variation below 1.2°C, while the aluminium components exhibited fluctuations exceeding ±6°C, particularly during periods of process instability. A documented production halt on day 2 corresponded with a 15–20% drop in recorded thermal intensity, highlighting the system's sensitivity to operational events. Although the system was not yet integrated with defect-tracking records, it effectively distinguished stable from unstable cooling conditions and identified temperature patterns indicative of potential quality issues. These results validate the system's potential as a diagnostic tool for monitoring thermal consistency and detecting anomalies in aluminium heat-treatment processes.

**Keywords:** Thermal Imaging · Quenching Process · Aluminum components · Real Time Monitoring · Manufacturing Efficiency · Heat Treatment.

## 1 Introduction

The production of high-quality aluminium components is essential in modern manufacturing, particularly in applications requiring precision, durability, and mechanical reliability. One such component is the machined aluminum piece, which significantly affects the performance and safety of the final product [3]. These components are typically produced by shaping aluminum bars through machining or forming, followed by quenching—a rapid cooling process critical for achieving strength, hardness, and wear resistance. However, consistent quality remains a challenge due to variability in the quenching process [2].

Quenching variability due to temperature swings, uneven cooling, or handling errors can cause defects like stress, distortion, or weakened properties [5, 7, 8]. Traditional quality checks, such as visual inspection and batch testing, are often imprecise, delayed, and subjective. To improve quality assurance, manufacturers are turning to advanced technologies that enable earlier, more objective detection of defects.

Advances in thermal imaging offer promising solutions for quality control in high-temperature manufacturing. Thermal cameras capture real-time surface temperature

changes, revealing key thermal dynamics during quenching [1]. Though established in other sectors, their real-time use in aluminum production remains limited. Studies show success in processes like casting and heat treatment [7,8], with applications in steel and automotive industries demonstrating thermal imaging's broader potential for detecting anomalies and improving material consistency [6].

Despite the advantages, several practical challenges remain in implementing thermal imaging technologies, particularly regarding system integration, real-time data processing, and accurate interpretation of thermal patterns. Additionally, linking temperature variations captured through thermal imaging with actual product quality requires advanced analytical approaches, such as deep neural network-based anomaly detection [4]. As a result, there is a clear need for a comprehensive thermal monitoring system that not only captures high-resolution temperature data but also analyzes it effectively to generate actionable insights for optimizing thermal treatment processes such as quenching.

This paper proposes such a system, utilizing thermal imaging to continuously monitor real-time temperature variations during the quenching of aluminum components. By delivering immediate feedback on thermal dynamics, the system enables proactive identification of potential issues and supports timely corrective actions. The paper outlines the system's design, implementation, and its implications for improving quality control, optimizing thermal processes, and enhancing overall production efficiency.

## 2 System Design and Architecture

The proposed thermal monitoring system enhances quality control in aluminum component production by continuously analyzing temperature distribution during the quenching process. Integrated into the existing manufacturing workflow, the system provides real-time feedback, enabling operators to detect and address potential defects before they impact product quality.

The system employs a FLIR ADK2.0 thermal camera with a 60 Hz frame rate and a 75° horizontal field of view to capture infrared radiation from the aluminum components and the quenching tank. This thermal data is transmitted to a processing unit via a data acquisition system, where it is stored and analyzed in real time. By comparing the recorded thermal profiles against predefined benchmarks, the system can identify temperature inconsistencies and cooling irregularities. When anomalies are detected, alerts are sent to operators through an interactive user interface that displays both live and recorded thermal data.

Fig. 1 illustrates the aluminum component production process, where manual inspection is performed at key stages such as shaping, heat treatment, surface finishing, and coating. Although these visual checks are important, they typically occur later in the workflow. As a result, defects introduced during earlier stages, particularly quenching, may go unnoticed until much later. By implementing direct thermal monitoring during quenching, it becomes possible to detect issues earlier. This real-time data, when linked with batch quality records, can significantly improve defect traceability and support overall process optimization.

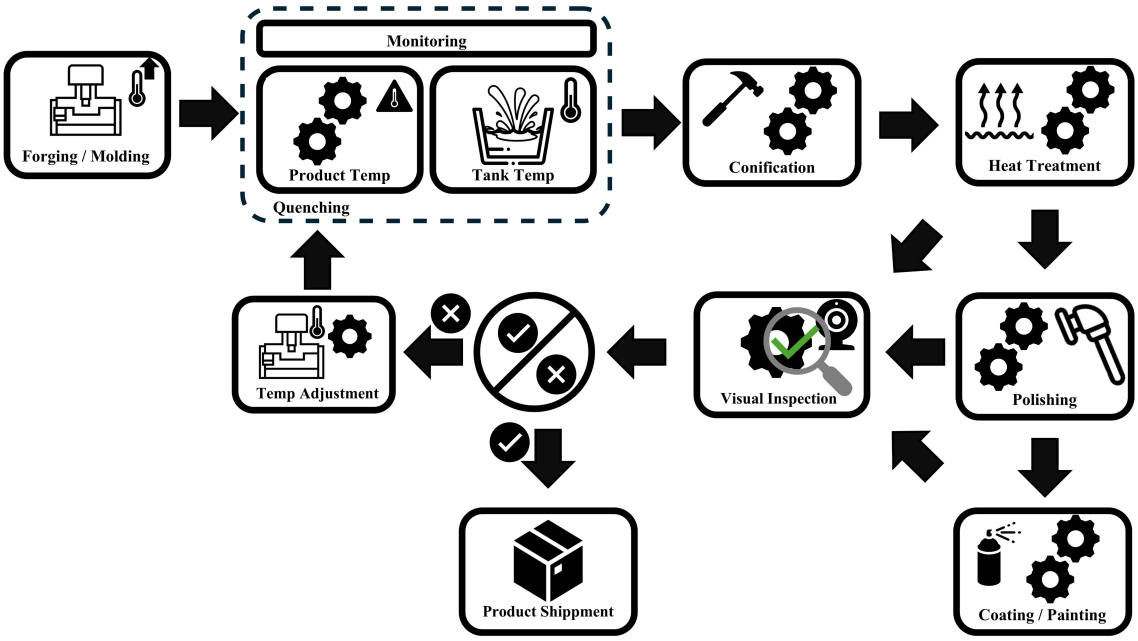



**Fig. 1.** Aluminum component production process, including forging, quenching with thermal monitoring, conification or shaping, heat treatment, surface finishing, quality inspection, and shipment.

Integrating thermal monitoring with production data creates a comprehensive dataset that links thermal signatures to documented defects. This enables the development of machine learning algorithms capable of evaluating component quality in real time. The system provides continuous, early detection of thermal inconsistencies during quenching, helping to maintain manufacturing standards, reduce waste, and improve overall efficiency.

## 3 System Overview

The monitoring system is designed to capture and analyze thermal data during the quenching stage of aluminum component production. Its goal is to detect thermal inconsistencies in real time, ensuring uniform cooling and minimizing defect propagation throughout the manufacturing process. The system consists of two primary hardware components:

- **FLIR ADK Thermal Camera** – Captures long-wave infrared (LWIR) radiation emitted from the aluminum surface, producing high-resolution temperature maps suitable for industrial monitoring. Key specifications are provided in Table 1.
- **ASUS Mini PC PN62S** – Acts as the local processing unit, receiving image data from the camera and handling data processing, anomaly detection, and communication with a central quality control server.

A Python-based algorithm was developed to process the incoming thermal images in real time. The algorithm segments the image into predefined regions of interest

**Table 1.** Key Specifications of the FLIR ADK Thermal Camera

| **Specification** | **Value** |
|---|---|
| Resolution | 640 × 512 pixels |
| Frame Rate | 30 Hz or 60 Hz |
| Thermal Sensitivity (NEdT) | < 50 mK |
| Spectral Band | 8–14 µm (LWIR) |
| Interface | USB 2.0, GMSL |
| Ingress Protection | IP67 (weather-resistant) |
| Operating Temperature Range | -40°C to +85°C |

and computes the thermal differential between these sections. This approach enables detection of localized cooling irregularities that might indicate process deviations or emerging defects. Each thermal frame is processed to extract temperature data, and these values are analyzed and stored for traceability.

The processed frames are also saved in a structured dataset along with metadata such as timestamp, region statistics, and process parameters. This growing dataset provides a foundation for training and validating machine learning models that can classify component quality, predict defect patterns, or recommend process adjustments. By automating the detection of thermal anomalies and systematically logging each instance, the system supports both short-term decision-making and long-term process optimization.

Initially, the system operates in a data acquisition mode, building a library of thermal signatures from components under varying conditions. For example, temperature deviations greater than ±5°C from a reference cooling curve within the first 10 seconds post-quench have been found to correlate with surface integrity issues. As operational knowledge accumulates, the system transitions toward real-time evaluation and automated alerting.

In its integrated form, the system provides immediate feedback during production. It enhances the consistency of quality control by supplementing visual inspections with precise thermal data. Fig. 2 illustrates the data flow and thermal monitoring integration points across the production process.

## 4 Prototype Implementation

The prototype implementation of the thermal monitoring system marked an initial step in assessing its effectiveness in improving quality control during aluminum component production. This phase involved the installation, calibration, and testing of the thermal camera to evaluate its ability to detect surface temperature inconsistencies and provide actionable feedback for process optimization.

The camera was positioned to capture real-time thermal variations during the transition from forming to quenching, with its angle optimized to fully cover the cooling phase (Fig. 3). Calibration accounted for emissivity, ambient reflections, and nearby heat sources to ensure temperature accuracy.

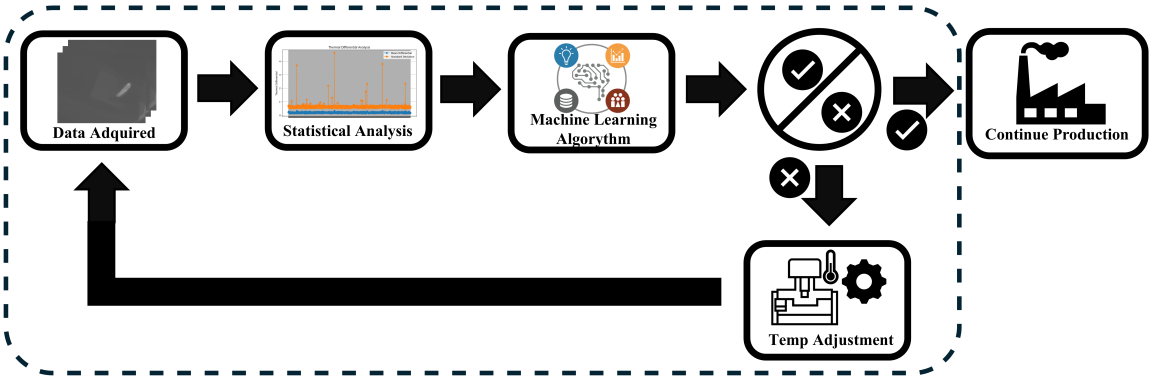



**Fig. 2.** Monitoring system functionality and data flow.

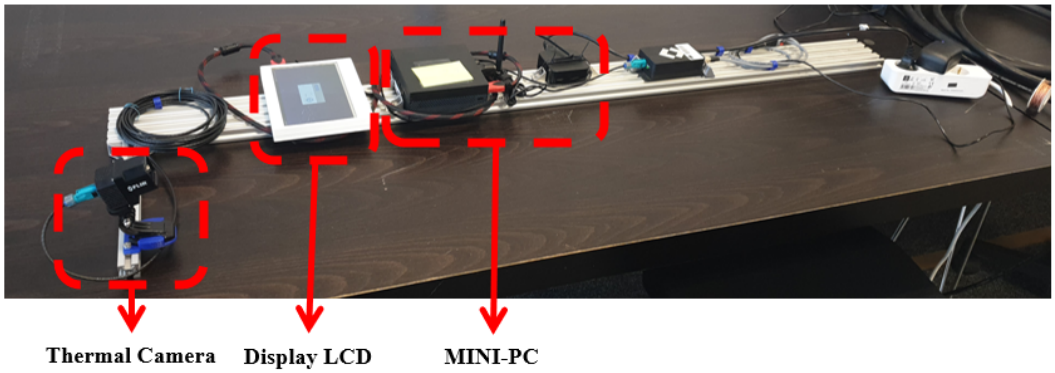



**Fig. 3.** Prototype composition and implementation.

A mini-PC continuously processed thermal frames, segmented regions of interest, and applied basic noise filtering. Tests under standard conditions confirmed stable frame rates, clear imaging, and proper synchronization with process timing—ensuring reliable data for analysis.

## 5 Experimental Results

The prototype thermal monitoring system successfully captured a large dataset of thermal images during the quenching process. It recorded surface temperature distributions of critical areas, including the alluminum produced component, the quenching fluid, and the entry plate that guides the piece into the tank. This dataset enables evaluation of thermal behavior across multiple production days and provides insight into cooling consistency.

Although not yet integrated with the factory's defect-logging system, the monitoring setup operated reliably and consistently detected temperature fluctuations. A structured dataset containing thousands of thermal frames was created, annotated with timestamps and segmented thermal regions. Preliminary

analysis was conducted to explore thermal consistency in relation to component positioning and production variability.

### 5.1 Thermal Distribution Across Production Days

Thermal data collected over four consecutive production days was analyzed to compare the distribution of temperatures in the monitored regions. Fig. 4 presents box plots summarizing the relative thermal distributions for the piece, fluid, and plate.

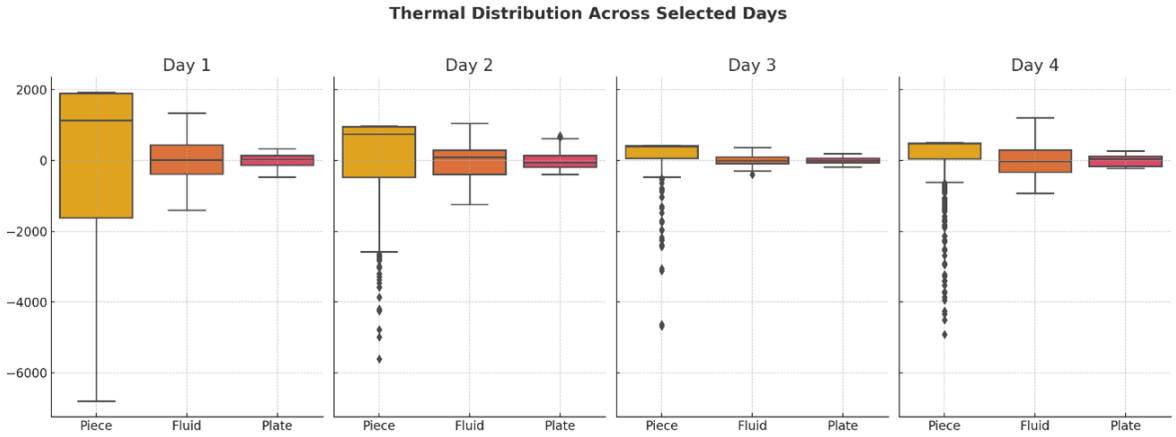



**Fig. 4.** Thermal distribution of Piece, Fluid, and Plate across four production days.

The fluid maintained stable thermal behavior across all days, reflecting its regulated role in the process. In contrast, the piece showed higher variability, particularly on days 2 and 3, while the plate exhibited increasing fluctuations on days 3 and 4. This suggests that the piece and plate are more affected by changes in process conditions.

### 5.2 Temporal Trends in Thermal Intensity

To investigate how temperature evolved during active production, thermal intensities were plotted over time for each monitored element. Fig. 5 displays time-series plots with mean-removed values across the four-day observation window.

Day 1 showed stable thermal signals with low fluctuation across all monitored areas. On day 2, increased variability in the piece's thermal profile was observed, particularly following a period of halted production. Days 3 and 4 exhibited higher-frequency fluctuations in both the piece and plate, with more pronounced spikes and dips. These shifts suggest the presence of non-uniform cooling behavior or possible disturbances in process stability.

Overall, the collected data confirmed that the thermal monitoring system is capable of capturing detailed temperature patterns that reflect real-time process dynamics. The distinction between stable elements (like the fluid) and variable ones (such as the piece and plate) reinforces the potential for further analysis and system refinement in later stages.

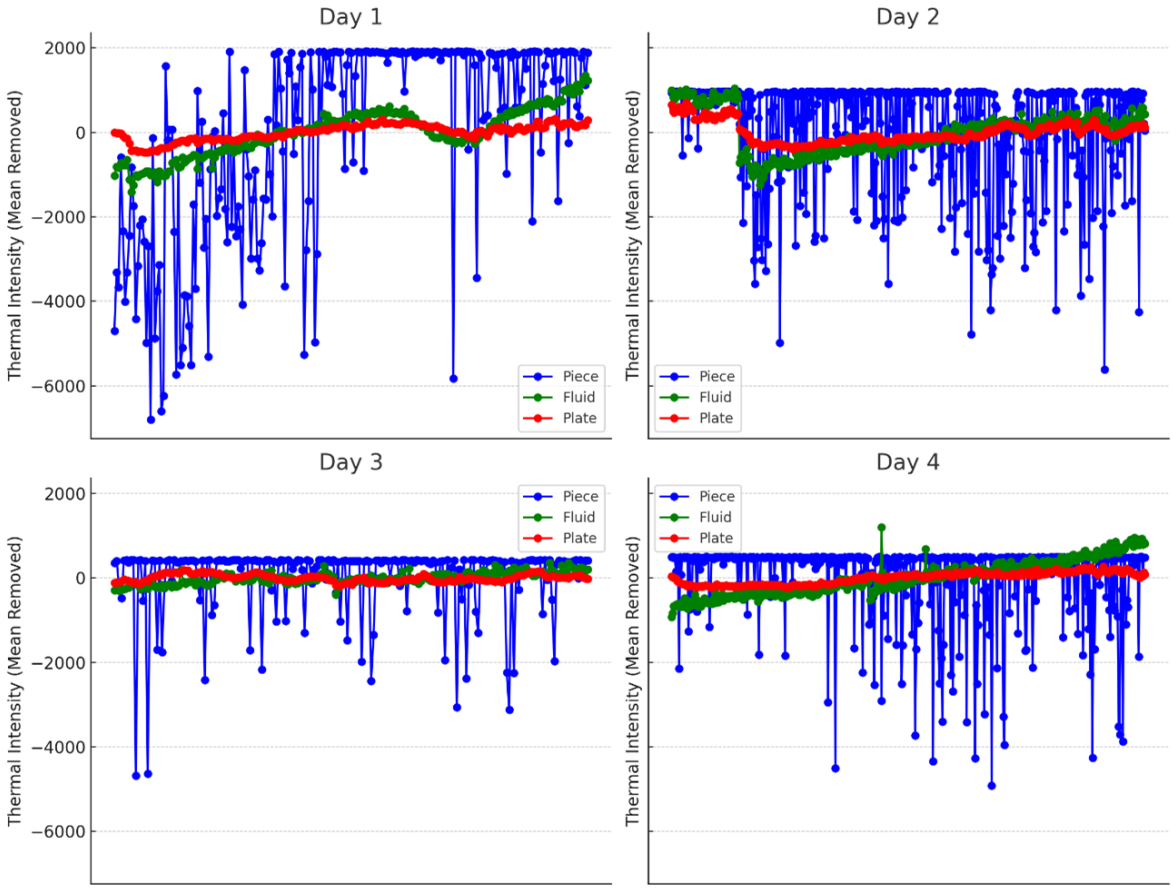



**Fig. 5.** Thermal intensity trends of Piece, Fluid, and Plate over four days. Values are relative and mean-normalized.

## 6 Analysis and Interpretation

The thermal monitoring system revealed distinct temperature behaviors across the component, quenching fluid, and entry plate. Although not yet linked to defect data, the system proved sensitive to process variations.

Box plots (Fig. 4) showed the fluid was consistently stable (SD $< 1.2$°C), while the component exhibited the highest variability (up to 4.8°C on Day 3, with outliers beyond $\pm 7$°C). The entry plate showed intermediate stability. These patterns highlight stronger thermal fluctuations at the component and quenching interface.

Time-series data (Fig. 5) supported these results. Day 1 showed stable behavior ($\pm 1.5$°C), while Days 3 and 4 had larger swings ($>6$°C), indicating process instability. On Day 2, a production halt (4:00–11:00 AM) caused a thermal drop and gradual post-restart recovery, confirming system responsiveness.

Overall, the component's wide variability and frequent outliers make it a strong indicator of cooling behavior. Even without defect integration, the system effectively distinguished stable from unstable conditions, supporting its role in monitoring quenching consistency.

## 7 Conclusion and Future Work

This study presented a thermal monitoring system for evaluating aluminum quenching. It provided continuous, non-contact thermal data, capturing temperature dynamics across components, entry plates, and fluid. While the fluid remained stable, significant surface variation on components demonstrated the system's ability to detect process inconsistencies.

Even without defect integration, the system identified operational shifts such as equipment changes and production halts, reinforcing its diagnostic value for thermal consistency and quality control.

Future work should focus on integrating thermal data with defect-tracking systems and conducting statistical validation against quality control outcomes. This advancement would enable a more robust assessment of the relationship between thermal behavior and final component integrity. These insights will also support ML-based defect prediction. Further work will enhance image robustness and expand camera coverage.

The system could be adapted to other industries for real-time feedback and improved process consistency.

## Acknowledgement

This research was funded by the "Agenda Mobilizadora para a inovação empresarial do setor das Duas Rodas (AM2R)" 01/C05-i01/2021 and LASI-LA/P/0104 /2020, co-funded from the "Mobilising Agendas/Alliances for Business Innovation" of the "Next Generation EU" program of Component 5 of the Portuguese Recovery and Resilience Plan (RRP), concerning "Investment and innovation", under the Regulation of the Incentive System for "Mobilising Agendas/Alliances for Business Innovation".

## References

1. Thermal imaging for quality control | 2019-12-02 | quality magazine, https://www.qualitymag.com/articles/95829-thermal-imaging-for-quality-control
2. Cao, P., Xie, G., Li, C., Zhu, D., Feng, D., Xiao, B., Zhao, C.: Investigation of the quenching sensitivity of the mechanical and corrosion properties of 7475 aluminum alloy. Metals **13**(10) (2023). https://doi.org/10.3390/met13101656, https://www.mdpi.com/2075-4701/13/10/1656
3. Hall, D.D., Mudawar, I.: Predicting the impact of quenching on mechanical properties of complex-shaped aluminum alloy parts. Journal of Heat Transfer **117**(2), 479–488 (05 1995). https://doi.org/10.1115/1.2822547, https://doi.org/10.1115/1.2822547
4. Lile, C., Yiqun, L.: Anomaly detection in thermal images using deep neural networks. Proceedings - International Conference on Image Processing, ICIP **2017-September**, 2299–2303 (7 2017). https://doi.org/10.1109/ICIP.2017.8296692
5. Ma, S., Maniruzzaman, M.D., MacKenzie, D.S., Sisson, R.D.: A methodology to predict the effects of quench rates on mechanical properties of cast aluminum alloys. Metallurgical and Materials Transactions B: Process Metallurgy and Materials Processing Science **38**, 583–589 (8 2007). https://doi.org/10.1007/S11663-007-9044-3/METRICS, https://link.springer.com/article/10.1007/s11663-007-9044-3
6. Tanner, D.A., Robinson, J.S.: Effect of precipitation during quenching on the mechanical properties of the aluminium alloy 7010 in the w-temper. Journal of Materials Processing Technology **153-154**, 998–1004 (11 2004). https://doi.org/10.1016/J.JMATPROTEC.2004.04.226

7. Zhang, L., Feng, X., Li, Z., Liu, C.: Fem simulation and experimental study on the quenching residual stress of aluminum alloy 2024. http://dx.doi.org/10.1177/0954405412465232 **227**, 954–964 (5 2013). https://doi.org/10.1177/0954405412465232, https://journals.sagepub.com/doi/abs/10.1177/0954405412465232
8. xun Zhang, Y., ping Yi, Y., quan Huang, S., Dong, F.: Influence of quenching cooling rate on residual stress and tensile properties of 2a14 aluminum alloy forgings. Materials Science and Engineering: A **674**, 658–665 (9 2016). https://doi.org/10.1016/J.MSEA.2016.08.017

# A Review of the Regulations Related to Digital Forensics of Mobile Devices

Carla Abreu Teixeira[1], Patrícia Anjos Azevedo[2], and Pedro Pinto[3]

[1] ESTG, Instituto Politécnico do Porto, Felgueiras, Portugal
8180690@estg.ipp.pt
[2] CIICESI, ESTG, Instituto Politécnico do Porto, Felgueiras, Portugal
pamv@estg.ipp.pt
[3] GECAD, Instituto Politécnico do Porto, IPP, Portugal
pfp@isep.ipp.pt

**Abstract.** Digital forensic analysis of mobile devices is a growing field in criminal investigations due to the volume and significance of stored data. In this context, this paper reviews the recent advances in digital forensics of mobile devices within the current regulations and highlights the associated gaps and challenges. This research is expected to contribute to assessing the current status and fostering the development of procedures for mobile device forensic analysis, aligned with national and European legal frameworks, including the Cybercrime Law and the General Data Protection Regulation (GDPR).

**Keywords:** Digital forensics · mobile devices · legal regulation · forensic procedure · data protection · encryption

## 1 Introduction

Mobile devices, such as smartphones and tablets, have become central to daily life, serving as large-scale data repositories for personal, professional, and criminal activities. Due to this capacity, they are considered critical sources of evidence in forensic investigations across various case types, including financial fraud, interpersonal violence, terrorism, and cybercrime [14]. A recent study[1] reports that 92.8% of the population uses mobile phones, with 99% of Portuguese users accessing the internet through such devices, averaging 3 hours and 35 minutes of daily use. This reinforces the relevance of mobile devices in the context of digital evidence collection.

However, mobile forensics presents significant challenges. Devices differ in architecture and security mechanisms, requiring diverse forensic techniques. Operating systems such as Android and iOS undergo frequent updates, complicating consistent data extraction. Additional barriers include encryption, biometric authentication, and PIN protection, which require continuous updates to forensic tools such as Cellebrite and Magnet AXIOM [1]. Legal and ethical challenges also emerge, particularly concerning the protection of personal data. The General Data Protection Regulation (GDPR) [5] mandates transparent and secure handling of personal information.

This study reviews the recent advances in digital forensics of mobile devices aligned with legal and regulatory standards, and highlights the associated gaps and challenges.

This effort intends to assess current and fragmented practices to propose a forensic procedure for mobile data acquisition.

[1] https://invoicexpress.com/relatorio-digital-portugal-2024/

## 2 Literature review

Digital forensics of mobile devices is an emerging and interdisciplinary field that integrates aspects of technology, law, ethics, and cybersecurity [3]. Thus, it is important to review the theories and technical advances related to this area, supporting theories that underpin the research and the legal and ethical implications in the context of digital forensics.

### 2.1 Digital Forensic Analysis Methodology

The methodologies associated with digital forensics aim to ensure the proper collection and interpretation of digital evidence, ensuring its validity in investigations and legal proceedings, as indicated by INTERPOL [7]. The specialized literature highlights an approach structured in three main stages: collection, preservation, and analysis, as recommended by NIST [11]. These stages were designed to minimize interference and protect the integrity of the evidence throughout the process.

The digital forensic process for mobile devices comprises three interdependent stages: data collection, preservation, and analysis. Data collection refers to the physical or logical extraction of information from mobile devices, using tools and techniques compatible with the specific hardware and operating system involved [2]. Preservation aims to maintain the integrity of the extracted data through the creation of forensic copies, the use of cryptographic hash functions, and the implementation of chain of custody protocols [6]. These practices seek to prevent data alteration and ensure legal admissibility. The analysis phase focuses on interpreting the preserved data to identify events relevant to an investigation. This stage includes tasks such as decrypting content, selecting pertinent information, and establishing relationships among data elements [7]. The integration of these three stages constitutes a procedural framework that supports the consistency and legal reliability of digital evidence throughout forensic investigations.

### 2.2 Digital Forensics of Mobile Devices

Digital forensics is a branch of forensic science that involves the recovery, preservation, and analysis of data from digital devices to use it as evidence in legal investigations. With the exponential increase in the use of mobile devices such as smartphones and tablets, these tools have become primary sources of digital evidence, containing critical data related to communications, locations, photos, videos, emails, and social media interactions. According to Kiran [10], the popularity of mobile devices and the growing dependence on them for everyday activities have turned them into essential elements in criminal investigations. A review of the literature on digital forensic analysis on mobile devices reveals a challenging scenario, characterized by rapid technological evolution, advanced security mechanisms, and the constant need to adapt forensic tools to changes in operating systems, such as Android and iOS [1]. The analysis of digital data stored on mobile devices requires the use of specialized techniques for data extraction and preservation, in addition to considering issues related to the integrity and admissibility of evidence in court.

Digital forensics of mobile devices faces significant technical challenges due to the dynamic and fragmented nature of mobile operating systems and the constant advances in security measures implemented by manufacturers. Data encryption, biometric authentication methods (such as fingerprints and facial recognition), the use of passwords and PINs, and the lack of a standardized methodology make the analysis more complex and prone to errors or failures in evidence collection [13]. According to Ramalho [15], the legal challenges are also notable. In many cases, breaches of mobile device security can be seen as a violation of user privacy, raising questions about the limits of digital investigation and the application of data protection legislation, such as the GDPR. Encryption and invasive analysis methods represent significant technical and legal challenges. For evidence to be admissible in court, the collection and analysis must respect the principles of legality, proportionality, and chain of custody, ensuring compliance with current standards.

## 3 Forensic Analysis - Legislation and Standardization

In this context, two interconnected theoretical areas can be devised: the theory of digital forensics and the theory of law, which deals with the admissibility and integrity of evidence. Digital forensics is based on a set of scientific methods and tools for the extraction, preservation, and analysis of digital data. The main theory in this review is the application of good practices and international standards, such as NIST publications [11], ISO/IEC 27043:2015 [6], and INTERPOL guidelines [7], which provide essential frameworks to ensure that the collection and analysis of digital evidence is carried out securely and reliably. The chain of custody model is a fundamental part of this theory, as it ensures the integrity of digital evidence, avoiding any contamination or manipulation that could compromise its admissibility in court [9]. Another essential theory is the theory of criminal procedural law, which deals with the admissibility and legal treatment of evidence. According to Portuguese legislation, such as the Code of Criminal Procedure [16] and the Cybercrime Law [12], digital evidence must be obtained legitimately, preserving the fundamental rights of individuals, such as the right to privacy. The GDPR also applies to the processing of personal data during forensic investigations. Establishes limits and principles that must be followed to ensure the legality, necessity, and proportionality of the processing of data; that is, the GDPR does not prohibit forensic investigations but imposes rules to ensure that the fundamental rights are respected during the process. The application of these legal standards in the context of digital forensics will be a central focus in this work, as there is an urgent need to align forensic analysis practices with legal requirements to ensure that the evidence extracted can be used effectively in court.

The literature discusses international standards that regulate digital forensic practices, including documents published by the National Institute of Standards and Technology (NIST), such as NIST Special Publication 800-101 [8]. This publication outlines specific procedures for conducting forensic analysis on mobile devices. These standards, widely adopted in jurisdictions such as the United States, influence forensic

methodologies and support the consistent collection and preservation of digital evidence following internationally recognized protocols.

In the Portuguese context, the literature identifies the absence of a unified national methodology for digital forensics. Although international guidelines exist, forensic practices in Portugal differ among investigative institutions, such as the Polícia Judiciária and the Public Prosecutor's Office. The lack of procedural standardization is presented as a significant challenge, highlighting the need for a consistent national framework to ensure the legal validity and uniform treatment of digital evidence.

## 4 Contributions and Gaps in Research

In the context of digital forensic analysis, it is crucial to ensure that digital evidence is both reliable and admissible in court, meeting all the necessary legal requirements. Without proper handling, even the most detailed investigation may fail if the evidence is deemed inadmissible. Therefore, the processes of accessing, collecting, preserving, and analyzing such evidence must follow established protocols that safeguard its authenticity, integrity, and legal compliance. This task requires professionals with the appropriate technical and scientific expertise to prevent any contamination that could undermine the credibility of the evidence [4]. The main gap is the lack of a standardized methodology for digital forensic analysis of mobile devices in Portugal. This lack of uniformity compromises the consistency of investigations.

## 5 Conclusions

This study examined the current state of digital forensic practices for mobile devices, identifying technical, legal, and procedural challenges. The analysis highlighted the fragmented nature of forensic procedures in Portugal and the absence of a standardized national framework. International guidelines, such as those from NIST and INTERPOL, were reviewed as references for best practices. Legal requirements, including those from the GDPR, the Portuguese Cybercrime Law, and the Code of Criminal Procedure, were considered to assess compliance in evidence handling. The findings underscore the necessity of developing a unified forensic methodology that aligns with legal standards and supports admissibility in court. Future efforts will focus on creating and testing a structured forensic procedure and a supporting digital tool to address these gaps.

Future work will focus on the development and validation of a structured forensic procedure for mobile device analysis. This procedure will include defined methods for data collection and preservation, such as the use of cryptographic hashes and chain of custody protocols, and data analysis, with attention to the reconstruction of investigative events and preservation of evidence integrity. In parallel, a digital tool will be designed to support the implementation of the procedure

## References

1. Mobile forensics tools and techniques for digital crime investigation: A comprehensive review. International Journal of Informatics and Communication Technology (IJ-ICT) **13**(2), 321–332 (Aug 2024). https://doi.org/10.11591/ijict.v13i2.pp321-332

2. Casey, E.: Digital Evidence and Computer Crime: Forensic Science, Computers and the Internet. Academic Press, 3rd edn. (2011)
3. Cavelty, M.D., Wenger, A.: Cybersecurity policy. In: Routledge eBooks. Routledge (2022). https://doi.org/10.4324/9781003110224, https://doi.org/10.4324/9781003110224
4. Dias, V.M.: The problem of cybercrime investigation. Data Vénia, Digital Law Journal **1**(01), 63–88 (July–December 2012), https://www.datavenia.pt/ficheiros/edicao01/datavenia01_p063-088.pdf
5. European Union: Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (general data protection regulation) (Apr 2016), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679, published in the Official Journal of the European Union L 119, 4 May 2016, p. 1–88
6. International Organization for Standardization: ISO/IEC 27043:2015 – Information Technology – Security Techniques – Incident Investigation Principles and Processes. https://www.iso.org/standard/44410.html (2015), accessed: 2025-04-30
7. INTERPOL: Guidelines for Digital Forensics First Responders. https://www.interpol.int/content/download/16243/file/Guidelines_to_Digital_Forensics_First_Responders_V7.pdf (2021), accessed: 2025-04-30
8. Kent, K., Chevalier, S., Grance, T., Dang, H.: Guide to Integrating Forensic Techniques into Incident Response: NIST Special Publication 800-101. Tech. Rep. 800-101, National Institute of Standards and Technology (2007), accessed: 2025-04-30
9. Khan, A.A., Uddin, M., Shaikh, A.A., Laghari, A.A., Rajput, A.E.: Mf-ledger: Blockchain hyperledger sawtooth-enabled novel and secure multimedia chain of custody forensic investigation architecture. IEEE Access **9**, 103637–103650 (2021). https://doi.org/10.1109/access.2021.3099037, https://doi.org/10.1109/access.2021.3099037
10. Kiran, S., Sanjana, J., Reddy, N.J.: Mobile phone addiction: Symptoms, impacts and causes — a review. In: International Conference on Trends in Industrial Value Engineering and Business and Social Innovation (ICTIVBSI). vol. 2019, pp. 81–86 (Mar 2019)
11. National Institute of Standards and Technology: Guidelines on cell phone forensics. NIST Special Publication 800-101 Rev. 1, U.S. Department of Commerce (2007). https://doi.org/10.6028/NIST.SP.800-101r1, https://doi.org/10.6028/NIST.SP.800-101r1
12. Portuguese Republic: Law no. 109/2009 of 15 september – cybercrime law. Official Gazette of the Portuguese Republic (2009), https://dre.pt/pesquisa/-/search/529821/details/maximized, establishes the legal framework applicable to cybercrime and the collection of electronic evidence
13. Raj, N.D.J.S.: A novel encryption and decryption of data using mobile cloud computing platform. IRO Journal on Sustainable Wireless Systems **2**(3), 118–122 (2021). https://doi.org/10.36548/jsws.2020.3.002, https://doi.org/10.36548/jsws.2020.3.002
14. Rakha, N.A.: Cybercrime and the law: Addressing the challenges of digital forensics in criminal investigations. Mexican Law Review pp. 23–54 (2024). https://doi.org/10.22201/iij.24485306e.2024.2.18892, https://doi.org/10.22201/iij.24485306e.2024.2.18892
15. Ramalho, D.S.: The use of malware as a means of obtaining evidence in criminal proceedings. Journal of Competition and Regulation **4**(16), 195–243 (2013)
16. República Portuguesa: Código de processo penal. https://diariodarepublica.pt/dr/legislacao-consolidada/decreto-lei/1987-34570075 (Feb 1987), decreto-Lei n.º 78/87, de 17 de fevereiro

# Intelligent Virtual Assistant for Industry 5.0

André Costa[1], Nuno Dinis[2], Luís Romero[1], and Pedro Miguel Faria[1]

[1] ADiT-Lab, Instituto Politécnico de Viana do Castelo, 4900-347 Viana do Castelo, Portugal
{afilipecosta, romero, pfaria}@estg.ipvc.pt
[2] RIOPELE, Pousada de Saramagos, Vila Nova de Famalicão, Portugal
nuno.dinis@riopele.pt

**Abstract.** Industry 5.0 introduces a new approach to industrial operations, emphasizing closer collaboration between humans and intelligent machines, promoting more resilient and sustainable production processes. This paper presents an Intelligent Virtual Assistant specifically designed for industrial environments, integrating advanced Natural Language Processing and machine learning techniques. The developed system features three main modes: maintenance, information, and prediction. Results from qualitative assessments with textile industry professionals revealed high usability and strong acceptance, highlighting the potential of IVAs to significantly enhance operational efficiency and intuitive human-machine interactions in line with Industry 5.0 objectives.

**Keywords:** Industry 5.0 · Intelligent Virtual Assistant · Machine Learning · Predictive Maintenance · Human-Machine Interaction

## 1 Introduction

### 1.1 Motivation

Industry 5.0 proposes a human–machine symbiosis in which skilled operators and data-driven systems work side-by-side to achieve resilient, sustainable and mass-customised production [11]. In high-mix textile manufacturing, for example, operators must juggle frequent product change-overs, stringent quality targets and tight delivery deadlines. While advanced automation hardware is commonplace, frontline personnel still spend a significant portion of their day *searching* for relevant knowledge (machine manuals, process parameters, historical alarms) or *escalating* issues to experts who are not always on-site. These frictions translate into longer mean-time-to-repair, higher scrap rates and, ultimately, reduced flexibility.

Intelligent Virtual Assistants (IVAs) promise to close that gap by providing a conversational layer – voice or chat – over heterogeneous plant data sources. Thanks to recent breakthroughs in Natural Language Processing (NLP) and large-scale representation learning, IVAs can now understand domain-specific jargon [4], contextualise follow-up questions [2], and generate actionable answers in real time. Our goal is therefore to investigate whether a *single* IVA, equipped with *maintenance*, *information* and *prediction* modes, can boost operator effectiveness while honouring the human-centric ethos of Industry 5.0.

### 1.2 Related Work

AI-enabled virtual assistants have delivered measurable benefits in consumer domains – e.g. retail chat-bots improve user experience and streamline logistics [17].

Transferring that success to industrial settings raises unique challenges: harsh acoustic conditions, proprietary terminology and the need to interface with Operational Technology (OT).

Early industrial prototypes relied on rule bases or FAQ retrieval. [1] coupled a rule-based chatbot with a digital twin to anticipate failures, reporting a 15 % reduction in unplanned downtime. The arrival of commercial Large-Language-Model (LLM) APIs has since sparked a new wave of assistants. Colabianchi *et al.* showed that an LLM-powered Digital Intelligent Assistant improved assembly quality and reduced operator workload in a controlled study [3]. At system level, Xia *et al.* integrated LLM agents with a modular production line, demonstrating autonomous task planning in a flexible cell [20]. Mazzei and colleagues focused on the I5.0 scenario, presenting an LLM-based assistant that bridges IoT data and managerial queries [5].

Surveys that take a broader view confirm the momentum: Li *et al.* catalogue more than forty GenAI use-cases across the manufacturing value-chain, yet identify a dearth of solutions that *simultaneously* cover maintenance support, live information access and predictive analytics [6]. Industrial suppliers echo that need; Siemens' *Industrial Copilot* positions generative AI as a shop-floor decision aid, although its current scope remains limited to single-asset diagnostics [18].

This paper addresses the above gap by introducing an IVA that unifies three capabilities within one architecture:

- Maintenance Mode:conversational troubleshooting grounded on equipment manuals;
- Information Mode: real-time querying of sensor and MES data;
- Prediction Mode:statistical forecasting of loom downtime to support proactive scheduling.

The system has been instantiated in a full-scale textile plant and evaluated through a mixed-methods usability study with expert operators. To the best of our knowledge, this is the first LLM-based assistant that concurrently addresses maintenance, information access and predictive analytics in a live production environment, thus advancing the Industry 5.0 agenda of collaborative, sustainable and human-centric manufacturing.

In contrast to existing approaches, our IVA uniquely integrates maintenance support, real-time information retrieval, and predictive analytics into a single conversational interface validated in a real industrial textile environment, directly addressing practical Industry 5.0 operational challenges.

## 2 Development of the IVA

The Intelligent Virtual Assistant (IVA) was developed with a modular architecture clearly divided into three main modes: Maintenance, Information, and Prediction. The system integrates Natural Language Processing (NLP), machine learning algorithms, and intuitive interfaces to effectively support operational needs in industrial contexts. Figure 1 presents a simplified diagram illustrating the architecture and the interaction between its modules and technologies.

In the **Maintenance Mode**, the IVA leverages technical documentation, extracting relevant information from industrial equipment manuals using the *pdfplumber* library

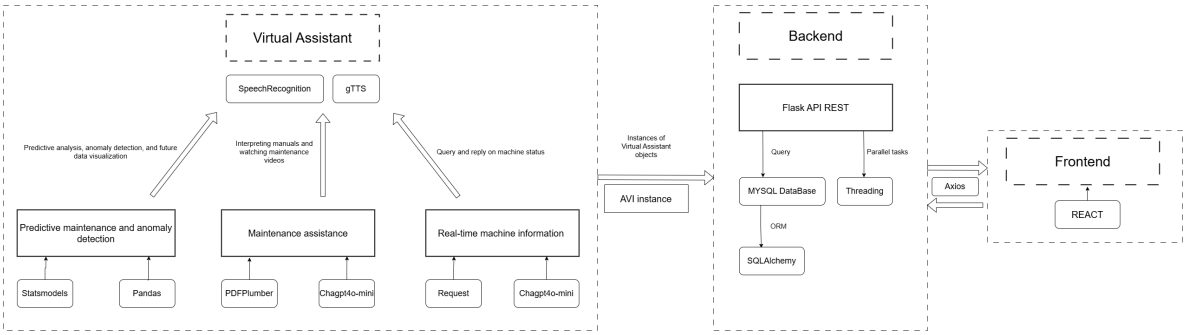



**Fig. 1.** General architecture diagram of the Intelligent Virtual Assistant (IVA)

[10]. This structured extraction allows operators to quickly solve maintenance issues. Additionally, this mode enables voice communication by integrating libraries such as *SpeechRecognition* [21] for voice recognition, *gTTS* [14] for speech synthesis, and *Pydub* [15] for audio processing.

The **Prediction Mode** incorporates predictive analytics through statistical modeling provided by the *Statsmodels* library [16], specifically utilizing the SARIMAX model to anticipate equipment malfunctions, enabling preventive measures and reducing downtime. Operational data management and preprocessing tasks are performed using the *Pandas* library [7].

In the **Information Mode**, the system communicates seamlessly with industrial APIs, employing the *OpenAI GPT-4o-mini* model [12] to interpret natural language queries from users. This ensures the provision of accurate, contextualized responses in real-time, enhancing decision-making processes in dynamic production environments.

The IVA's backend was implemented using *Flask* [13], coupled with *SQLAlchemy* [19] for robust database interactions and session management. This setup allows efficient asynchronous communication with the frontend interface, developed in *React* [9], selected for its modular architecture and performance efficiency suited to industrial operational contexts.

## 3 Results

The evaluation of the IVA prototype was conducted with 10 textile industry professionals, including operators and maintenance technicians, each with a minimum of five years of relevant experience. The methodology combined quantitative and qualitative approaches. The quantitative evaluation utilized the System Usability Scale (SUS) [8], resulting in an excellent score of 85 points, indicating very high usability.

Qualitative data were gathered through semi-structured interviews to provide deeper insights into the user experience. Participants unanimously described the IVA as intuitive, practical, and easily integrated into their daily operational routines, particularly praising the Maintenance Mode for its simplicity and effectiveness.

Participants also suggested several specific enhancements:

– Improve usability across the different operational modes.

– Provide detailed comparative data visualization.
– Optimize the database structure to facilitate faster access to operational information.

Regarding the Information Mode, participants noted that NLP integration effectively provided accurate and contextualized real-time responses. The responses were largely satisfactory, with minor suggestions for further refinement to enhance contextual awareness.

## 4 Conclusion

This paper presents the development of an Intelligent Virtual Assistant designed to meet the operational and technical challenges of Industry 5.0. The proposed solution combines advanced Natural Language Processing technologies, machine learning and voice-text interaction to offer intelligent and contextualized support to maintenance and operation technicians. The modular architecture, made up of a backend, frontend and intelligent core, enabled effective integration with real industrial systems, guaranteeing scalability, adaptability and usability.

The practical validation carried out in a textile factory demonstrated the concrete applicability of the solution, with high acceptance from end users. The ease of use, the clarity of the responses and the effectiveness of the operating modes - maintenance, prediction and information - in solving technical problems and anticipating faults stood out. The qualitative analysis also revealed a short learning curve, good integration with the industrial context and a highly satisfactory user experience, with a SUS score of 85 points.

However, opportunities for improvement were also identified, particularly in the predictability of the data and the structuring of the information presented. Based on these observations, the following vectors for future development of the AVI are outlined:

– Improving Interaction with AI: Making conversation more natural and proactive, allowing the assistant to recognize intentions implicit in commands and anticipate user needs with contextual suggestions.
– Knowledge Base Enrichment: Expand the range of technical data available (detailed manuals, historical logs, advanced specifications), enabling more comprehensive and analytical responses, including equipment comparisons and optimization recommendations.
– Prediction of Future Faults: Integrate analysis of technical reports and maintenance history with real-time data to detect fault patterns and issue early warnings, increasing the system's proactivity.
– Voice recording of maintenance tasks: Allow technicians to record interventions by dictation, with automatic transcription and intelligent data validation, reducing administrative effort and improving the traceability of operations.

These advances aim to strengthen the role of AVI as a strategic tool in supporting the operation of smart factories, promoting closer collaboration between humans and

intelligent systems, and consolidating the Industry 5.0 paradigm. The future implementation of these functionalities will not only increase the assistant's effectiveness but also expand its contribution to increasingly demanding, dynamic and human-centered industrial environments.

## Acknowledgement

This work received financial support from the integrated project Pacto Mobilizador TEXP@CT – Pacto de Inovação para a Digitalização dos Têxteis e do Vestuário (TC-C12-i01, Bioeconomia Sustentável n.º 02/C12-i01/202), promoted by the Recovery and Resilience Plan (PRR), Next Generation EU, for the period 2021-2026.

## References

1. Barbosa, A.d.S., Silva, F.P., Crestani, L.R.d.S., Otto, R.B.: Virtual assistant to real time training on industrial environment. In: Peruzzini, M. (ed.) Transdisciplinary Engineering Methods for Social Innovation of Industry 4.0. pp. 33–42. IOS Press (2018). https://doi.org/10.3233/978-1-61499-898-3-33
2. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. In: Advances in Neural Information Processing Systems 33. pp. 1877–1901 (2020), https://arxiv.org/abs/2005.14165
3. Colabianchi, S., Costantino, F., Sabetta, N.: Assessment of a large language model based digital intelligent assistant in assembly manufacturing. Computers in Industry **162**, 104129 (2024). https://doi.org/10.1016/j.compind.2024.104129, https://doi.org/10.1016/j.compind.2024.104129
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. pp. 4171–4186 (2019), https://aclanthology.org/N19-1423
5. Figliè, R., Turchi, T., Baldi, G., Mazzei, D.: Towards an llm-based intelligent assistant for industry 5.0. In: Proceedings of the 1st International Workshop on Designing and Building Hybrid Human–AI Systems (SYNERGY 2024). CEUR Workshop Proceedings, vol. 3701 (2024), https://ceur-ws.org/Vol-3701/paper7.pdf
6. Li, Y., Zhao, H., Jiang, H., Pan, Y., Liu, Z., Wu, Z., Shu, P., Tian, J., Yang, T., Xu, S., Lyu, Y., Blenk, P., Pence, J., Rupram, J., Banu, E., Liu, N., Wang, L., Song, W., Zhai, X., Song, K., Zhu, D., Li, B., Wang, X., Liu, T.: Large language models for manufacturing. arXiv preprint arXiv:2410.21418 (2024), https://arxiv.org/abs/2410.21418
7. McKinney, W.: Pandas: Biblioteca para análise e manipulação de dados em python (2010), pyData. Disponível em: https://pandas.pydata.org/. Acessado em 6 de fevereiro de 2025
8. MeasuringU: Recent advances with the system usability scale. MeasuringU Blog (2017), https://measuringu.com/sus-advances/
9. Meta Platforms, I.: React – a javascript library for building user interfaces (2025), https://react.dev/
10. Miller, J.: pdfplumber: Biblioteca para extração de texto estruturado de documentos pdf (2020), mIT License. Disponível em: https://github.com/jsvine/pdfplumber. Acessado em 6 de fevereiro de 2025
11. Nahavandi, S.: Industry 5.0—a human-centric solution. Sustainability **11**(16) (2019). https://doi.org/10.3390/su11164371, https://www.mdpi.com/2071-1050/11/16/4371
12. OpenAI: Openai api: Modelos de linguagem avançados para análise de texto (2024), openAI Documentation. Disponível em: https://platform.openai.com/docs/api-reference. Acessado em 6 de fevereiro de 2025
13. Projects, P.: Flask documentation (2025), disponível em: https://flask.palletsprojects.com/en/stable/patterns/sqlalchemy/. Acessado em 7 de fevereiro de 2025

14. Rijsdijk, T.: gtts: Conversão de texto para fala utilizando a api do google (2019), gTTS Documentation. Disponível em: https://gtts.readthedocs.io/en/latest/. Acessado em 6 de fevereiro de 2025
15. Robert, J.: Pydub: Manipulação simples de áudio com python (2019), pydub Documentation. Disponível em: https://github.com/jiaaro/pydub. Acessado em 24 de fevereiro de 2025
16. Seabold, S., Perktold, J.: statsmodels: Econometric and statistical modeling with python. In: 9th Python in Science Conference (2010)
17. Seranmadevi, R., Chakraverty, S., Raj, B., Kudapa, V.K., Hepziba, R.E., Suleimenova, K.: Utilisation of virtual assistant and its impact on retail industry. In: 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS). pp. 1729–1733 (2022). https://doi.org/10.1109/ICICCS53718.2022.9788243
18. Siemens AG: Scaling roll-out of generative ai with siemens industrial copilot. Press release (May 2024), `https://press.siemens.com/global/en/pressrelease/siemens-xcelerator-scaling-roll-out-generative-ai-siemens-industrial-copilot`, accessed: 2025-06-15
19. SQLAlchemy: Sqlalchemy documentation (2025), disponível em: `https://docs.sqlalchemy.org/13/orm/contextual.html`.
20. Xia, Y., Shenoy, M., Jazdi, N., Weyrich, M.: Towards autonomous system: Flexible modular production system enhanced with large language model agents. arXiv preprint arXiv:2304.14721 (2023), `https://arxiv.org/abs/2304.14721`
21. Zhang, A.: Speechrecognition: Biblioteca para reconhecimento de fala em python (2017), speech Recognition (Version 3.11) [Software]. Disponível em: https://pypi.org/project/SpeechRecognition/. Acessado em 6 de fevereiro de 2025

# Defining Requirements for Post-Stroke Hand Rehabilitation Devices

Fernando Rocha[1,2] and Fernando Veloso[1,2]

[1] 2Ai – School of Technology, IPCA, Barcelos, Portugal
[2] LASI, Associate Laboratory of Intelligent Systems, Guimarães, Portugal
{frocha, fveloso}@ipca.pt

**Abstract.** Loss of hand movement following a stroke severely compromises the quality of life of patients, requiring continuous care by professionals and caregivers to provide functional recovery. However, this process can place pressure on both therapists and patients, mainly due to constant supervision and attendance at therapeutic sessions. In this context, assistive technologies have been integrated to enhance the rehabilitation process by enabling functional training, as well as monitoring patient progress. In this study, we have conducted a literature review to better understand user requirements for the development of such technologies. The study was carried out across three distinct databases, encompassing 17 included studies. The study's findings are delineated by 45 requirements, subsequently organized into five groups, according to their nature.

**Keywords:** Stroke · Hand Rehabilitation · User Requirements.

## 1 Introduction

Stroke is one of the leading causes of death, and a major agent of disability worldwide. Consequently, even if patients can overcome the condition, numerous unpredictable complications may still lead to a decrease in quality of life [1]. Considering the essential role played by the hand in daily life, loss of movement after a stroke severely compromises the patient's degree of independence, preventing them from performing different tasks that require precise coordination of the limbs [2]. Despite the apparent incapacity, the patient may experience frustration with their condition. This frustration can lead to the development of psychological complications, such as depression and anxiety [3]. Furthermore, the rehabilitation process is a demanding phase in post-stroke recovery, where continuous and thorough monitoring by the rehabilitation team, as well as any entities close to the patient involved in therapy, is crucial [1]. Over the years, specialized assistive technologies have been developed to provide a more dynamic and interactive answer to traditional therapy, by reducing the burden on therapists and caregivers, while maintaining intensive treatment [4]. Building upon previous studies, this paper aims to understand the most fundamental design requirements expressed by patients and therapists, as the articulation of user needs is imperative for the conception and long-term therapeutic viability of novel assistive technologies. To achieve this, a workflow was designed to ensure proper organization of the research (Fig. 1). Our workflow encompasses four main milestones, beginning with the definition of the scope of the study and systematically organizing the breadth of the research, followed by reviewing the literature and analyzing the findings for thematic commonalities. Future work outlines further crucial steps to be taken after the review.

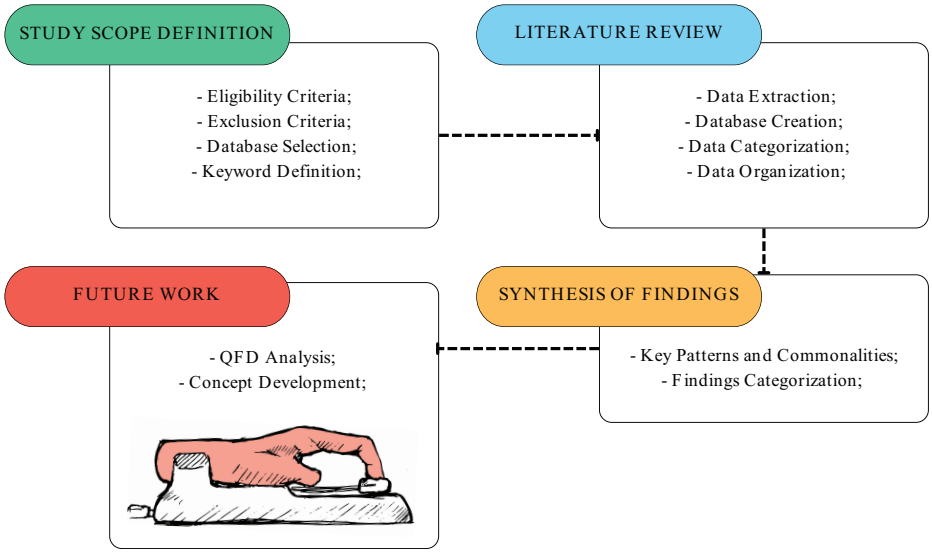



**Fig. 1.** Workflow outlining all key milestones within the study.

## 2 Methods

### 2.1 Databases

The study was conducted through three different databases, with a time frame between 2015-2025: IEEE, PubMed, and Web of Science, respectively, by using the following search string: (1) IEEE: Stroke; Hand Rehabilitation; User Requirements; (2) PubMed/Web Of Science: Stroke; hand rehabilitation; user needs.

### 2.2 Inclusion and Exclusion Criteria

This paper only included studies published in peer-reviewed journals and conference proceedings. Studies that emphasized interviews or other direct interactions with therapists and/or patients were included. To delimit the focus on hand rehabilitation, articles highlighting elbow or shoulder rehabilitation were excluded; Bearing in mind other impairment-inducing pathologies, such as Spinal Cord Injury (SCI), and other phenomena such as aging and assorted hand injuries, are distinct from stroke in the way they manifest impairment, only stroke-related studies were considered. Articles focusing on software and user interfaces were excluded. Similarly, studies that did not include any discernible requirements were also excluded.

## 3 Results

After carefully analyzing the retrieved records, 17 studies were included in this study. After the full reading process, the included studies were compiled into a structured database to extract the most relevant requirements. 45 requirements were identified from the literature and subsequently grouped into five categories, as illustrated by Fig. 2, through an inductive thematic analysis. Previous research on identifying user requirements has exposed various approaches to categorizing findings, resulting in inconsistent classifications. To address this issue, we have synthesized the findings of these studies into a comprehensible classification system, enabling a more cohesive analysis of different requirements and their implications for the overall design of the device.

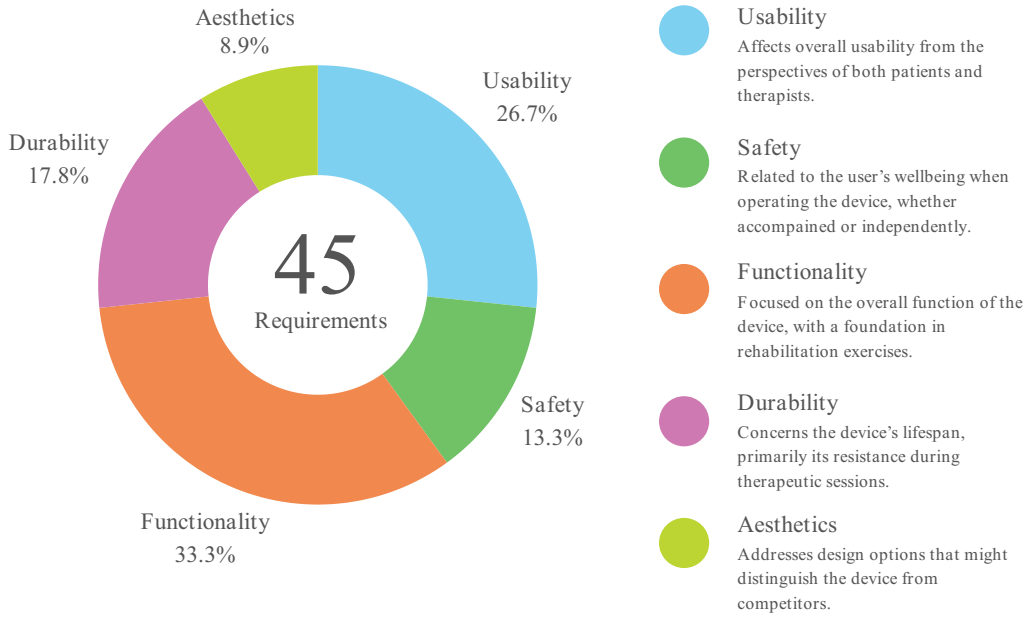



**Fig. 2.** Distribution of identified user requirements across the five identified domains.

## 4 Discussion

The identified requirements underline the need for devices to serve as an effective link between patients and therapists. Undoubtedly, safety-driven requirements should be prioritized when developing these devices, as patients will rely on them during their therapy routine [5]. Lightweight design, comfort when worn, smooth surfaces, and overall safety in using the device without the help of caregivers constitute the most frequently identified safety-related requirements. Similarly, requirements such as ease of donning and doffing the device, ease of setting up, and adaptability to various

hand dimensions can also contribute to the resulting safety of the device. Regarding the usability and functionality of the device, it has been identified that the ability to provide feedback to users, track/measure hand activity, and subsequently provide such information to caregivers is a highly desirable feature [5, 6]. Regarding therapeutic goals, the ability of the device to be used to train specific grasp movements, be compatible with activities of daily living (ADL), and provide adjustable resistance to better suit the degree of patient impairment has the potential to leverage rehabilitation progress, especially if the device can be used at home [7]. Other relevant requirements, such as long battery life, resistance to water, easy to clean, compact dimensions, and implementation of gamification strategies, have also been identified, and are equally significant for the conception of long-lasting, portable, and captivating devices [5, 8, 9]. There is a clear recognition that although user requirements are identifiable, several tend to be subjective to a certain extent. In this context, some authors emphasize tangible metrics that can be considered to meet specific requirements. Boser et al. conducted an exploratory study involving six clinicians and eight stroke patients, expressing a consensus regarding an acceptable maximum weight of 200g, a maximum volume of 5 x 5 x 3 $cm^3$, and the ability to use the device ranging from 6 hours to a full day, denoting the need for a long battery life to meet user demands [10]. Radder et al. identified through their study involving 28 participants, among elderly, stroke patients, and therapists, that a set of movements, including spherical grasp, palmar grasp, lateral grasp, and key pinch, are among the most important movements in need of support. Furthermore, the authors define a quick initialization time between <1 and 2 minutes, as well as a battery that can last for a full day, intermittently, without charging the device [8]. However, when considering adapting such findings to one's own project, these must be subject to careful discussion alongside therapists and end-users, as the typology of devices, as well as end-users, can differ significantly.

## 5 Future Work and Conclusion

As a next step, the implementation of a quality function deployment (QFD) approach will facilitate the translation of the identified requirements, particularly those of a more subjective nature, into a comprehensible and organized hierarchy for the development of a new device. This process will ensure that each requirement is assigned a relative weight, thereby enhancing decision making during the design process [11]. In our study, we have identified user requirements through a structured review of the literature for the development of post-stroke hand rehabilitation devices. Our findings highlight a wide range of valuable requirements that can be studied and implemented for the conception of novel devices. In fact, such requirements play a vital role during the design and development phases of novel devices, as the demand for technological innovations in the field of rehabilitation increases.

## Acknowledgement

This work is funded by the HfFP - Health From Portugal Innovation Pact (Notice no. 01/C05-i01/2021, Mobilizing Agendas/Alliances for Reindustrialization RE-C05-i01.01 and Green Agendas/Alliances for Business Innovation RE-C05-i01. 02) within the scope and co-financing of the "Mobilizing Agendas for Business Innovation" of the "Next Generation EU" programme of Component 5 of the Recovery and Resilience Plan (RRP), relating to "Capitalization and Business Innovation", under the "Agendas for Business Innovation" Incentive System Regulation.

## References

1. S. Chohan, P. Venkatesh, and C. How, "Long-term complications of stroke and secondary prevention: an overview for primary care physicians," *Singapore Medical Journal*, vol. 60, pp. 616–620, 12 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7911065/
2. F. Mawase, K. Cherry-Allen, J. Xu, M. Anaya, S. Uehara, and P. Celnik, "Pushing the rehabilitation boundaries: Hand motor impairment can be reduced in chronic stroke," *Neurorehabilitation and Neural Repair*, vol. 34, pp. 733–745, 08 2020.
3. L. Liu, M. Xu, I. J. Marshall, C. D. Wolfe, Y. Wang, and M. D. O'Connell, "Prevalence and natural history of depression after stroke: A systematic review and meta-analysis of observational studies," *PLOS Medicine*, vol. 20, p. e1004200, 03 2023. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC10047522/
4. W. H. Chang and Y.-H. Kim, "Robot-assisted therapy in stroke rehabilitation," *Journal of Stroke*, vol. 15, p. 174, 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3859002/
5. C.-H. Lo, Y.-C. Ko, and H.-P. Chiu, "User needs for a robotic hand rehabilitation device in home-based therapy following a stroke: A user-centered approach," *Sensors and Materials*, vol. 36, p. 1731, 05 2024.
6. A. L. van Ommeren, L. C. Smulders, G. B. Prange-Lasonder, J. H. Buurke, P. H. Veltink, and J. S. Rietman, "Assistive technology for the upper extremities after stroke: Systematic review of users' needs," *JMIR Rehabilitation and Assistive Technologies*, vol. 5, p. e10510, 11 2018.
7. G. Prange, L. Smulders, J. van Wijngaarden, G. Lijbers, S. Nijenhuis, P. Veltink, J. Buurke, and A. Stienen, "User requirements for assistance of the supporting hand in bimanual daily activities via a robotic glove for severely affected stroke patients," pp. 357–361, 08 2015.
8. B. Radder, A. Kottink, v. , D. Oosting, J. Buurke, S. M. Nijenhuis, G. B. Prange, and J. S. Rietman, "User-centred input for a wearable soft-robotic glove supporting hand function in daily life," 08 2015.
9. S. Forbrigger, V. DePaul, T. C. Davies, E. Morin, and K. Hashtrudi-Zaad, "Home-based upper limb stroke rehabilitation mechatronics: challenges and opportunities," *Biomedical Engineering Online*, vol. 22, 07 2023.
10. Q. A. Boser, M. R. Dawson, J. S. Schofield, G. Y. Dziwenko, and J. S. Hebert, "Defining the design requirements for an assistive powered hand exoskeleton: A pilot explorative interview study and case series," *Prosthetics and Orthotics International*, p. 030936462096394, 10 2020.
11. D. Wang, H. Yu, J. Wu, Q. Meng, and Q. Lin, "Integrating fuzzy based qfd and ahp for the design and implementation of a hand training device," *Journal of Intelligent & Fuzzy Systems*, vol. 36, no. 4, pp. 3317–3331, 2019.

# System for Detecting Rubber Imperfections in Extrusion Lines

Paulo Sousa[1], Pedro Carneiro[2], and José Henrique Brito[1]

[1] 2Ai – School of Technology, IPCA, Barcelos, Portugal
a26834@alunos.ipca.pt,jbrito@ipca.pt
[2] CMIP, Continental Mabor, Lousado, Portugal
pedro.ricardo.carneiro@conti.de

**Abstract.** During the production process of rubber components for tyre manufacturing, is subject to stringent quality standards, particularly during the extrusion stages of the process, where imperfections have the potential to compromise the integrity of the final product. It is imperative to implement continuous visual inspection during these stages to ensure product compliance. The presence of elevated technical requirements often complicates the identification of non-conformities during the manufacturing process, with these discrepancies often becoming apparent only at later stages of production. This limitation can result in the discarding of raw materials, which can compromise operational efficiency and process sustainability.
To address these limitations, this research proposes the development and implementation of an visual inspection system based on computer vision and artificial intelligence. The model that was selected was YOLOv8 due to its efficiency in real-time object detection, combined with its high accuracy and low computational requirements. The objective of this system is to detect surface defects in rubber directly on the extrusion line, thereby ensuring immediate action. The experimental results obtained demonstrate that the system achieved mAP50 of 99% and mAP50:95 of 70%, along with an average inference rate of 5 milliseconds per image. These metrics serve to substantiate the viability of the proposed solution.

**Keywords:** Defect Detection · Computer Vision · Artificial Intelligence · Rubber.

## 1 Introduction

In the realm of vehicle safety, tires play a critical role as the sole point of contact between vehicles and the road surface. The World Health Organization reports that approximately 40% [1] of traffic accidents are related to tires.
The tire industry currently faces challenges in its quality control procedures, particularly in visual inspection automation. The existing defect assessment methods are characterized by reduced efficiency and intense workload, falling short of meeting the demands associated with large-scale production of high-quality products. The current approach to checking surfaces for nonconformities is inadequately developed, lacking a dedicated system capable of effectively identifying irregularities.
This late detection is often due to inadequate fusion of components and is exacerbated by challenging conditions such as continuous production flow and poor lighting, particularly when dealing with dark-colored components.
This article describes a deep learning methodology for the effective recognition of irregularities detected directly on the production line. The application in question was implemented using the YOLOv8 algorithm, which is widely used in object detection problems [7], [2]. A data set of grayscale images with defects was created and annotated with defect detection bounding boxes. The images were initially

automatically weakly annotated and the annotations were manually corrected using Label Studio [10]. The aim of the project is to extend current studies on the subject of object detection, adapting them to the requirements of extrusion lines.
The article is organized into the following sections. In Section 2, a comprehensive review of the existing literature is provided, with a focus on the rationales that underpin the methodological framework used in this study. Section 3 provides a comprehensive overview of the architectural design of the proposed system, which includes its hardware and software components. In Section 4, the dataset utilized is characterized, the training parameters adopted are specified, and the results obtained are presented. These results are based on a rigorous evaluation of the model's performance. Finally, Section 5 offers a comprehensive summary of the primary contributions of the work, together with an in-depth discussion of the impact and potential of the developed system in the context of its applications.

## 2 Related work

With advances in deep learning techniques, new methodologies have emerged that make use of neural networks. On the work of [6] , the study outlines a machine vision system for inspecting rubber rollers, using both conventional and deep learning approaches. The classification of defects and the determination of their positions are carried out using the YOLOv7 methodology, underlining the superior accuracy and robustness of the proposed method.While the traditional method faces challenges in categorizing defects, the deep learning approach, despite the need for labeled data and efficient hardware, reveals extraordinary accuracy in classification. The balanced combination of these methods promises to optimize classification accuracy and efficiency. Future improvements are aimed at reducing computational costs to enable wider application in industry.

A substantial example is the work by [2], which investigates methods for monitoring cracks in wind turbine blades, integrating classification, detection, segmentation, and evaluation of the level of failure in a single artificial intelligence model. This multifaceted approach highlights the potential of neural networks in complex applications. Using an improved algorithm based on YOLOv8, called "Multivariate Information Perception You Look Only Once" (MIP-YOLO), the study demonstrates how the combination of advanced techniques can improve accuracy and efficiency in crack detection, contributing to the stability and economic benefits of wind energy.

TraCon [3], for real-time detection of traffic cones in road environments using deep learning techniques. Despite substantial progress in the realm of object detection within road environments, the prevailing focus of research has been on vehicles and pedestrians, thereby neglecting objects pertinent to road management and maintenance, such as traffic cones. The present study utilizes the YOLOv5 algorithm, a well-regarded approach that has demonstrated notable efficiency and speed in object detection tasks. The proposed method was applied to a set of RGB images of road works, collected from various sources. The findings indicate a high degree of detection

accuracy, with an IoU (Intersection over Union) score reaching as high as 91.31%. The article underscores the significance of traffic cone detection for enhancing road safety and the efficiency of road construction.

The present article [7] proposes a generalized model for real-time detection of flying objects using the YOLOv8 architecture, which is regarded as state-of-the-art in single-shot object detectors. The work addresses the challenges inherent in detecting flying objects, such as the wide variation in sizes, speed rates, occlusions, and complex backgrounds. The generalized model was trained on a dataset comprising 40 classes of flying objects, thereby enabling the extraction of abstract feature representations. Subsequently, transfer learning was applied with the parameters learned on a dataset more representative of real environments, generating a refined model. The generalized model demonstrated a mAP50 of 79.2% and mAP50-95 of 68.5%, with an average inference speed of 50 frames per second in 1080p videos. The refined model exhibited no loss in inference speed and demonstrated a substantial enhancement in accuracy, attaining mAP50 of 99.1% and mAP50-95 of 83.5%. The article under scrutiny emphasizes the efficacy of YOLOv8 in addressing both small and moving objects, underscoring its capacity for expeditious and precise real-time applications.

In study [13], a comprehensive dataset was developed that encompasses all categories of tire defects. We introduced a new framework called DCScNet to perform tire defect classification. Compared to leading classification algorithms, DCScNet learned features in a fully unsupervised way, layer by layer, achieving a remarkable classification accuracy of 96.8%.

Convolutional Neural Network (CNN) [12] with an attention mechanism was proposed to identify foreign objects in the raw materials on a conveyor belt during mining production. This approach is based on the concept of an attention mechanism and the combination of semantic segmentation. Attention modules were integrated into the CNN to focus on the features of the highlighted region and inhibit irrelevant background, resulting in an effective improvement in detection accuracy. The proposed method showed demonstrated improvements in foreign object detection results in a production context, using the visualization technique. The results showed that the method correctly identified 97% of foreign objects in 1871 sets of test images. The Mean Jaccard Index (MIOU) of the optimal model reached 91.24%, and the inference speed was over 15 fps, meeting real-time requirements.

In addition, the study by [8] proposed an innovative method for detecting defects in resin films, using a combination of computer vision techniques, and supervised neural networks. This method makes use of a large field-of-view microscope with 2K resolution to capture high-quality images. Using an adjusted LeNet-5 network model, the method locates and predicts defect types based on shape. In addition, deep autoencoders are used to reconstruct images of normal samples, identifying anomalies by comparing the original image with the reconstructed one. This approach demonstrated high accuracy in locating and inspecting defects in resin films. Experimental results indicated that the method can effectively identify microscopic defects, highlighting the robustness and accuracy of the model in identifying anomalies. This study underlines the importance of integrating deep learning techniques with conventional methodologies has led to a

substantial reduction in the necessity for manual inspections. This development presents a considerable opportunity for implementation in a variety of industrial contexts.

To resolve the problem presented and the research carried out, the YOLOv8 model was selected for implementation of the object detector. This model is based on the PyTorch framework. The model architecture is made up of three parts: Backbone employs a sophisticated convolutional neural network (CNN), which has been designed to extract multi-scale features from input images. This backbone, which may be an advanced version of CSPDarknet [11] or another efficient architecture, captures hierarchical feature maps that represent the textures, which are crucial for accurate object detection. The backbone has been optimized for both speed and accuracy, incorporating depth-separable convolutions or other efficient layers to minimize computational overhead while retaining representational power. The neck serves to refine and merge the multiscale features extracted by the backbone. The model utilizes an optimized version of the Path Aggregation Network (PANnet) [5] and the Feature Pyramid Network (FPN) [4], which facilitates enhanced information flow across various feature levels. This multiscale integration is imperative for the detection of objects of varying sizes and scales. It is probable that YOLOv8 will incorporate alterations to the original PANet, in order to further optimize memory usage and computational efficiency. The head is responsible for generating the final predictions, which include the coordinates of the bounding boxes, object confidence scores and class labels. The model introduces an anchor-free method [9] which has been demonstrated to simplify the prediction process, reduce the number of hyperparameters, and improve the model's adaptability.

## 3 System Architecture

The system presented uses a dataset of images of extrusion lines to identify imperfections. It is imperative to note that each image captured by an RGB camera is processed to grayscale and then used in a script that has been developed specifically to label it accurately. This is done to meet the criteria that have been established to be considered as an imperfect. The implementation of Label Studio [10] facilitates the execution of additional adjustments, thus improving the quality of the labeling process. As imperfections are not a constant occurrence, a script was developed to expand the image database by rotating the images 180 degrees. It is important to note that no other data augmentation methods are employed, since, due to the stability of the detection zone, only rotation is considered relevant. The annotations are generated in .txt files, which store information about the classes and bounding boxes annotated. The model was trained using an NVIDIA RTX 4000 ADA with 16 GB of memory. The integration of these images into the YOLOv8 algorithm is seamless, ensuring optimal performance. The efficacy of this system is predicated on its capacity to categorise and localise objects in accordance with the loss function, thereby converting the target detection problem into a regression problem [2] [7]. This algorithm incorporates the most advanced detection technologies available at the time and optimises its implementation for best practices [6].

**Fig. 1.** Annotation image

## 4 Experiential Development

**Dataset and Training** The data set under consideration consists of 6,700 images, which were divided into three distinct sets. The first set, constituting 60% of the data, was allocated for training purposes. The second set, comprising 20% of the data, was designated for validation. The third set, which consists of an equal proportion of images as the validation set, was allocated for testing purposes. It should be noted that this dataset also includes unlabelled images without defects (empty .txt files). The model was trained with batches of 16 for 200 epochs. The input images are of dimensions 4096x2 pixels, which are then reshaped to 640x640x3 pixels when fed to in the model. To avoid the problem of overfitting, an early stopping strategy with a patience of 50 was employed. This approach therefore stopped the training process when the model's validation loss did not improve.

**Evaluation Metrics** The performance of the algorithm is evaluated through the application of several analytical methodologies. The precision (P), recovery (R), mean average precision (mAP), and F1 methods are utilized to assess the efficacy of the algorithm.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F1 = \left( \frac{2}{\text{Recall}^{-1} + \text{Precision}^{-1}} \right) 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

$$mAP = \frac{1}{S} \sum_{j=1}^{S} AP(j) \tag{4}$$

True positive (TP) is defined as the number of positive samples that have been correctly predicted. False positive (FP) is the number of positive samples that have

been incorrectly predicted and false negative (FN) is the number of negative samples that have been incorrectly predicted.

**Results** In Fig. 2, the training parameters and graphs associated with the learning process employed in the training are represented. The evaluation of the model performance is based on quantitative metrics extracted during training, such as box, objectivity, classification, box val, objectivity val and classification val. For these parameters, lower values are indicative of reduced loss, thus resulting in a more suitable fit for the data model. Concurrently, metrics such as Precision, Recall, mAP@0.5 and mAP@0.5:0.95 are utilised as direct indicators of the model's capacity to recognise and locate objects. In this case, higher values are indicative of superior performance. Consequently, the synergy of minimal losses in the validation and training parameters, concomitant with elevated mAP values, engenders an efficient model, characterised by high discriminative capacity and robustness in the object detection task.

**Table 1.** Model performance on validation data.

| P | R | F1 | mAP@0.5 | mAP@0.5:0.95 | PR |
|---|---|---|---|---|---|
| 0.98 | 0.99 | 0.98 | 0.99 | 0.70 | 0.99 |

The test dataset comprises a total of 1340 images, each of which may contain multiple defects. See detail in Table 1. Subsequent to the implementation of the neural network on the test dataset, the model demonstrated an aptitude for accurately identifying 1096 defects. However, certain instances of defects eluded detection, a fact that was reflected in the performance results obtained. The performance indicators for the model's training and validation iterations must be considered in order to ensure the validity of the results obtained. A consistent downward trend is evident in the losses incurred during training and validation, thus indicating that the model has successfully learned to represent the data and minimise predictive errors. The loss values have been found to stabilise in recent epochs, thus avoiding the presence of any obvious signs of overfitting. The decrease in box loss indicates an improvement in the accuracy of the bounding boxes, while the reduction in cls loss reflects an increase in the reliability of the classification of detected objects. With regard to performance metrics, the precision and mAP50 values demonstrate an upward trend throughout the training process, attaining elevated values. This verifies the model's considerable capacity to accurately identify pertinent objects, exhibiting a minimal false positive rate. The mean average precision (mAP) at multiple IoU thresholds also demonstrates a progressive enhancement, attaining a value close to 0.70, which signifies a robust capacity to detect objects at varying scales and positions. A slight decrease in recall has been observed in the final epochs, which may indicate a tendency towards greater selectivity on the part of the model, favouring precision over sensitivity, and potentially resulting in the omission of some true

positives. This behaviour must be considered when adjusting confidence thresholds or balancing precision and recall. It is important to note that the model has an average inference time per image of 5 milliseconds, which confirms its high computational efficiency and reinforces its viability for real-time applications.
In summary, the results obtained demonstrate the effectiveness of the model in terms of both accuracy and efficiency, validating its applicability in industrial environments where high levels of performance and real-time response are required.

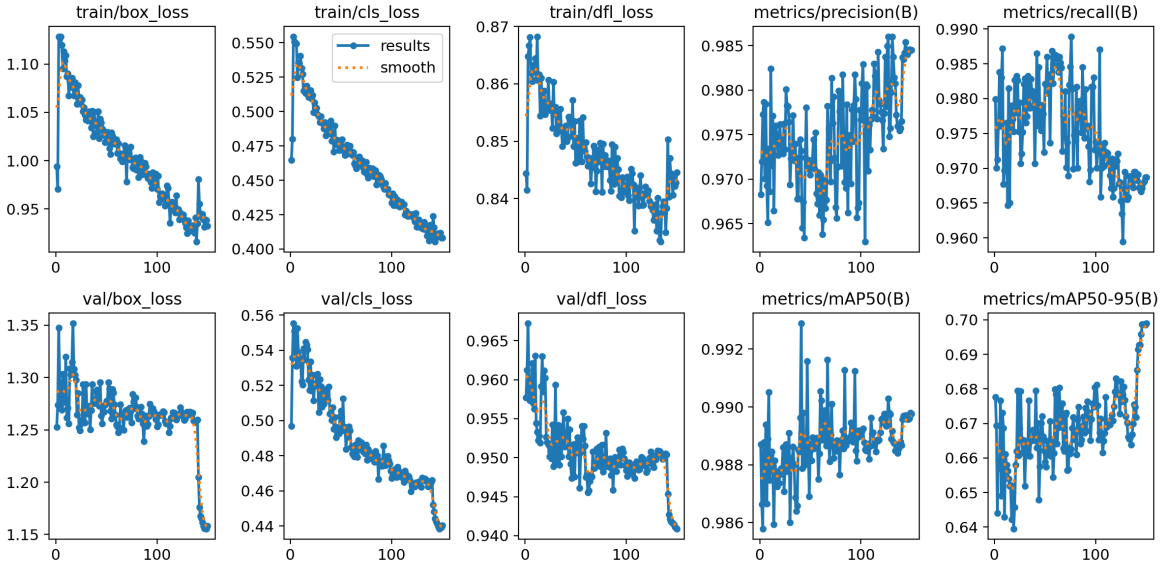



**Fig. 2.** Training graphs

## 5 Conclusions

The present article focuses on the development of a computer vision system for real automated inspection of rubber in extrusion lines. The approach based on deep neural networks, namely through the YOLOv8 architecture, has proven to be effective in the simultaneous detection and localization of product faults. The model facilitates the identification of defects in real time, determining their precise locations and affected areas, thereby enabling an accurate assessment of the product's compliance with quality requirements. The implementation of this methodology signifies a substantial advancement in the domain of automatic quality control within challenging industrial environments, thereby fostering enhanced efficiency, reliability and traceability in real time inspection process.

## References


1. Erdogan, S.: Explorative spatial analysis of traffic accident statistics and road mortality among the provinces of turkey. Journal of Safety Research **40**(5), 341–351 (2009). https://doi.org/https://doi.org/10.1016/j.jsr.2009.07.006
2. Hang, X., Zhu, X., Gao, X., Wang, Y., Liu, L.: Study on crack monitoring method of wind turbine blade based on ai model: Integration of classification, detection, segmentation and fault level evaluation. Renewable Energy **224**, 120152 (2024). https://doi.org/https://doi.org/10.1016/j.renene.2024.120152, https://www.sciencedirect.com/science/article/pii/S0960148124002179
3. Katsamenis, I., Karolou, E.E., Davradou, A., Protopapadakis, E., Doulamis, A., Doulamis, N., Kalogeras, D.: Tracon: A novel dataset for real-time traffic cones detection using deep learning (2022), https://arxiv.org/abs/2205.11830
4. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection (2017), https://arxiv.org/abs/1612.03144
5. Liu, S., Qi, L., Qin, H., Shi, J., Jia, J.: Path aggregation network for instance segmentation (2018), https://arxiv.org/abs/1803.01534
6. Nguyen, T.H., Nguyen, H.L., Bui, N.T., Bui, T.H., Vu, V.B., Duong, H.N., Hoang, H.H.: Vision-based system for black rubber roller surface inspection. Applied Sciences **13**(15) (2023). https://doi.org/10.3390/app13158999, https://www.mdpi.com/2076-3417/13/15/8999
7. Reis, D., Kupec, J., Hong, J., Daoudi, A.: Real-time flying object detection with yolov8 (2024), https://arxiv.org/abs/2305.09972
8. Sheu, R.K., Teng, Y.H., Tseng, C.H., Chen, L.C.: Apparatus and method of defect detection for resin films. Applied Sciences **10**(4) (2020). https://doi.org/10.3390/app10041206, https://www.mdpi.com/2076-3417/10/4/1206
9. Solawetz, J.: What are anchor boxes in object detection? Roboflow (2020), https://blog.roboflow.com/what-is-an-anchor-box/
10. Tkachenko, M., Malyuk, M., Holmanyuk, A., Liubimov, N.: Label Studio: Data labeling software (2020-2025), https://github.com/HumanSignal/label-studio, open source software available from https://github.com/HumanSignal/label-studio
11. Wang, C.Y., Liao, H.Y.M., Yeh, I.H., Wu, Y.H., Chen, P.Y., Hsieh, J.W.: Cspnet: A new backbone that can enhance learning capability of cnn (2019), https://arxiv.org/abs/1911.11929
12. Zhang, K., Wang, W., Lv, Z., Fan, Y., Song, Y.: Computer vision detection of foreign objects in coal processing using attention cnn. Engineering Applications of Artificial Intelligence **102**, 104242 (2021). https://doi.org/https://doi.org/10.1016/j.engappai.2021.104242, https://www.sciencedirect.com/science/article/pii/S0952197621000890
13. Zheng, Z., Shen, J., Shao, Y., Zhang, J., Tian, C., Yu, B., Zhang, Y.: Tire defect classification using a deep convolutional sparse-coding network. Measurement Science and Technology **32**(5), 055401 (mar 2021). https://doi.org/10.1088/1361-6501/abddf3, https://dx.doi.org/10.1088/1361-6501/abddf3

# Increasing Warehouse Efficiency Using Quality Tools In Factory Operations

Zied Ben Cheikh[1], Artur Rossi[2], Jose Barbosa[1], and Paulo Leitão[1]

[1] Instituto Politécnico de Bragança, Bragança, Portugal
[2] Research Center in Digitalization and Intelligent Robotics (CeDRI), Laboratório Associado para a Sustentabilidade e Tecnologia em Regiões de Montanha (SusTEC), Instituto Politécnico de Bragança, Bragança, Portugal
{cheikh, jbarbosa, pleitao}@ipb.pt
[3] EORE Research Group, Federal University of Technology-Paraná, UTFPR, Brazil
arthurhgrossi@hotmail.com

**Abstract.** Managing the storage is becoming increasingly complicated to assess, with rising costs and more complex inventories in small and medium-size enterprises (SMEs). Effective storage management necessitates accurate tracking, reduction of excess inventory and rapid access to materials. Analysis of storage methods is essential for continuous improvement to increase operational efficiency, cost management and productivity. This paper explains how to better understand warehouse problems through careful analysis and application of Ishikawa diagrams. It facilitates the setting of specific improvement objectives by investigating every issue and determining the source of the problems in the warehouse. The study also talks about using simple technology to make factories work better. It aims to increase the efficiency of the work inside the warehouse, using smart objectives to help for drawing up an action plan for improvements with statistics results and how to integrate the new technologies of industry 4.0 in this methodology.

**Keywords:** Ishikawa · Storage Management · SMART goals · PDCA.

## 1 Introduction

The Toyota Productive System (TPS) is one of the most recognized production systems in the world, serving as a model for many of the successful production methods in the industry, especially for those who are looking for adequacy in the practice of real case studies. As a strong competitive advantage, the TPS offers its customers, through its production process, the full satisfaction of their expectations, and simultaneously, obtaining economic benefits surpassed only by companies with similar productive qualities [1].

Storage management analysis is very important. It helps companies make better use of space, reduce costs, and keep track of inventory efficiently. Good warehouse management makes it easier to find and use materials quickly. This reduces wastes and improves the fluidity of operations. Regular review of storage management enables companies to improve performance and remain competitive. This approach makes it possible to achieve the key objectives of the optimized production, to ensure the company's continued growth and to enable it to respond to the evolution of the market [2].

The application of push system instead of the pull system led to overproduction, which in turn ended on a overstocking in the warehouse. It causes difficulties because of

the small amount of warehouse space, having trouble doing tasks quickly as a result of it. To adhere to the First In, First Out (FIFO) principle and maintain the warehouse's organization, it must constantly move containers around. Not only that, but the lack of effective systems makes it easy to misidentify and use the wrong materials. This disrupts the work and may cause mistakes when orders are sent out. To analyse and identify problems, a methodology based on ishikawa approach is developed where 3 layer approach is presented to analyse environment and measurement, machine and material and manpower and methods groups of root problems.

The rest of the paper is organized as follows: Section 2 reviews analysis methods for diagnosing warehouse inefficiency. Section 3 outlines the methodologies and redesigned tools used in the case study. Section 4 presents the company and the structure of the warehouse where it based this case study, mentions the problem description, the application of ishikawa analysis on the anomalies to help for defining the objective goals, showing the improvement solutions and the results of the analysis and the objective goals drawn in this case with a statistics results analysis. Section 5 discuss long-term effectiveness the benefits of this methods comparing the existing ones and how to improve this methodology by implementing the new technologies of industry 4.0. Finally, section 6 concludes the paper, limitations and identifies areas for future work.

## 2 Warehouse inefficiency analysis: A state of the art

Warehouse inefficiencies could be analyzed by different types of problems solving methods. These methods helps decision maker to well understand the root causes and to fix solutions for improvement.

### 2.1 5W2H method

An initial approach to explain the issue (the problem, the failure) is the 5W2H method (Why, Where, Who, What, When, How and How much). It is used for finding the underlying cause of a problem or system failure. 5W2H tool were used as decision-making guidelines. It helps company in the process of creating manufacturing techniques. This approach is crucial to learning how to solve problems. It is also included in the first TPS (Toyota Production System) training. Toyota Company's scientific approach is based on TPS, according to its originator, Taichii Ohno. In order to find remedial and preventive measures, the five Why inquiries help identify the root of the issue. The tool has been expanded outside Toyota and is used now in all the sectors [3]. Not just determining the cause of an issue, 5W2H aims to assist in resolving it and preventing it from occurring in the future. A corporation can take corrective action to prevent recurrence of problems by determining the reasons behind failures [4].

### 2.2 Pareto analysis

A Pareto chart is essentially a frequency block diagram in statistics that shows the relative frequency of several attributes in descending order. Making this classification is

a necessary first step before implementing the differentiation and allocation correction procedures. While many researchers have documented instances in which Pareto charts demonstrate the validity of the 80/20 principle, in actual practice, some classifications result in Pareto charts that are less helpful for managers because the relative frequency corresponding to 20% of the attributes is significantly less than 80%, making it difficult to concentrate on a small number of attributes [5].

### 2.3 5 Whys

Using a five-question "why" technique, 5-Whys Analysis seeks to systematically identify the root causes of issues. It has been frequently used in many industrial contexts to explore issues relating to production, quality, reliability, safety, and environment. Its origins back to the Toyota Production System. It is easier to develop and put into action appropriate countermeasures when the technique allows for the distinction between causal variables and root causes. When included into operational tools, 5-Whys Analysis can produce measurable and practical results. But doing so necessitates having a strong understanding of both the issue being looked into and the system being studied. Effective use of the method can reduce these disadvantages and make sure it is sufficient for problem solving, even though it might not always indicate common causes or provide repeatability of results [6].

### 2.4 PDCA cycle

Four phases are repeated as part of this procedure to control and enhance the supply chain management process or the company's habits. To put it another way, this approach consists of four stages that are used to monitor any deviations and make necessary adjustments in order to improve business operations. The Planning, Conducting, Testing, and Implementation steps also referred to as the Deming Phase mark the conclusion of the PDCA process. The plan do check act cycle was created by Deming as a four-step methodical approach to issue solving [7]. Several corrective, temporary, and permanent actions are generated by the PDCA cycle idea. The goal of continuous and corrective action is to eliminate the root cause. Quick steps to address and resolve issues [8].

## 3 A new methodology for warehouse inefficiency analysis

This approach, presented in Fig. 1, aims to have comprehensive analysis by identifying and categorizing causes of problem, a clear visual representation for the relation between the causes and effects, a flexible variation of problem, making easy the brainstorming sessions, finding useful insights and rank causes by importance and using an easy design that simplify the process.

The approach follows four PDCA-aligned steps—detect anomalies via team input, identify root causes using Ishikawa diagrams, define SMART improvement objectives, and implement a targeted action plan—integrating analysis and goal-setting to enhance efficiency.

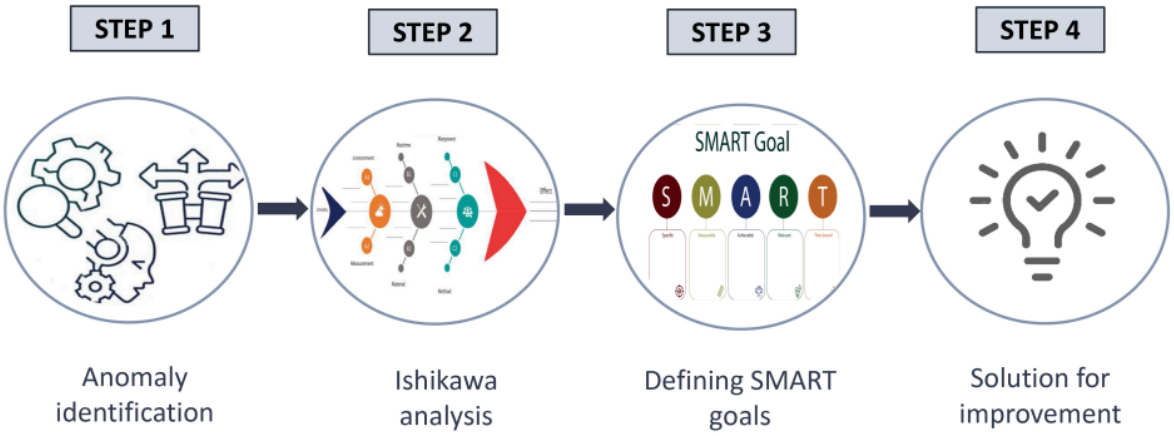



**Fig. 1.** Methodology developed

## 4 Application in a small push storage warehouse

### 4.1 Company delineation

The factory produces cold-formed parts using industrial stamping processes with modern equipment and technologies, ensuring autonomy and meeting group needs. After parts are removed from the finished goods area, they are packed in boxes and prepared for shipping, with loads moved to trucks using forklifts. A larger forklift operates externally for safety and space reasons, while a smaller one moves boxes inside. The warehouse is divided into functional areas: the center stores finished parts, the top side houses compressors, coils, and inspection areas, the right side handles loading/unloading and low-rotation items, and the bottom side manages packaging and shipment preparation for efficient dispatch.

### 4.2 Features for the methodology to address

**Anomaly identification** This step leads for well understanding and analyzing this challenges.

The identified challenges are the limited Space Impacts, Organizational Challenges, Lack of Visual Management, Transport Challenges,Security Considerations.

**Ishikawa analysis** In Fig. 2 , an Ishikawa diagram that depicts causes and effects of the problem of limited space anomaly is shown.

Moving for drawing smart goals to effectively address the identified issues.

**Defining SMART goals** The objectives smart outlined on the Fig. 3.

This objective focuses on improving operational efficiency by reducing container movements by 20% within six months, leading to cost savings, streamlined processes, and reduce wastes.

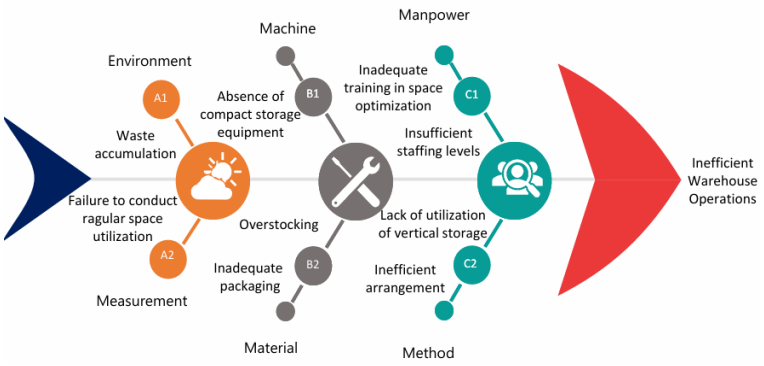



**Fig. 2.** Analysis of limited space challenge

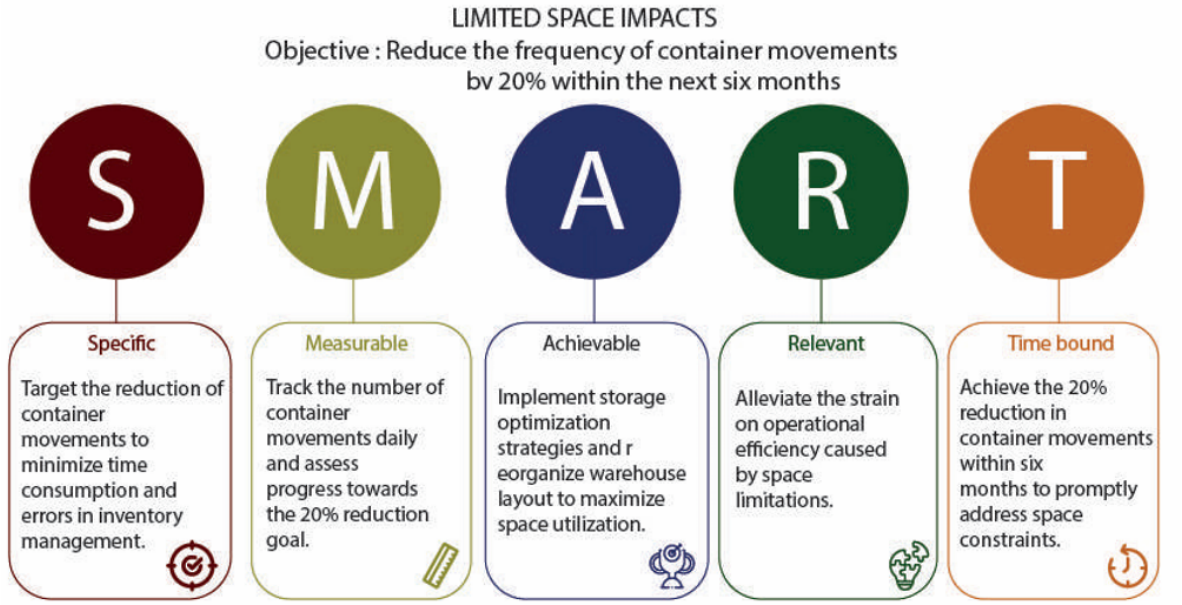



**Fig. 3.** SMART goal of limited space

**Improvement solutions** The previous sections explained how to perform the diagnostic of the problems encountered using the Ishikawa diagram. During this analysis, it turns out that all the problems have common points which is the lack of visual management that impact the organization and the process approach inside the warehouse. This causes problems like moving containers around and making mistakes in finding and picking items, the waste of time during this unjustified movement also for finding the container wished and accidents that may occur due to insufficient awareness signs.

To fix these problems, some solutions for improvement are outlined in the Fig. 4.

The provision of implementing this methodology enabled conducting operations in a more orderly, effective and error-free way. This in turn built the capacity of the organisation to find out the key anomalies and their causes. The methodology was able through broad and systematic approach to achieve significant operating improvement in five key areas: space constraint, organization problems and absence

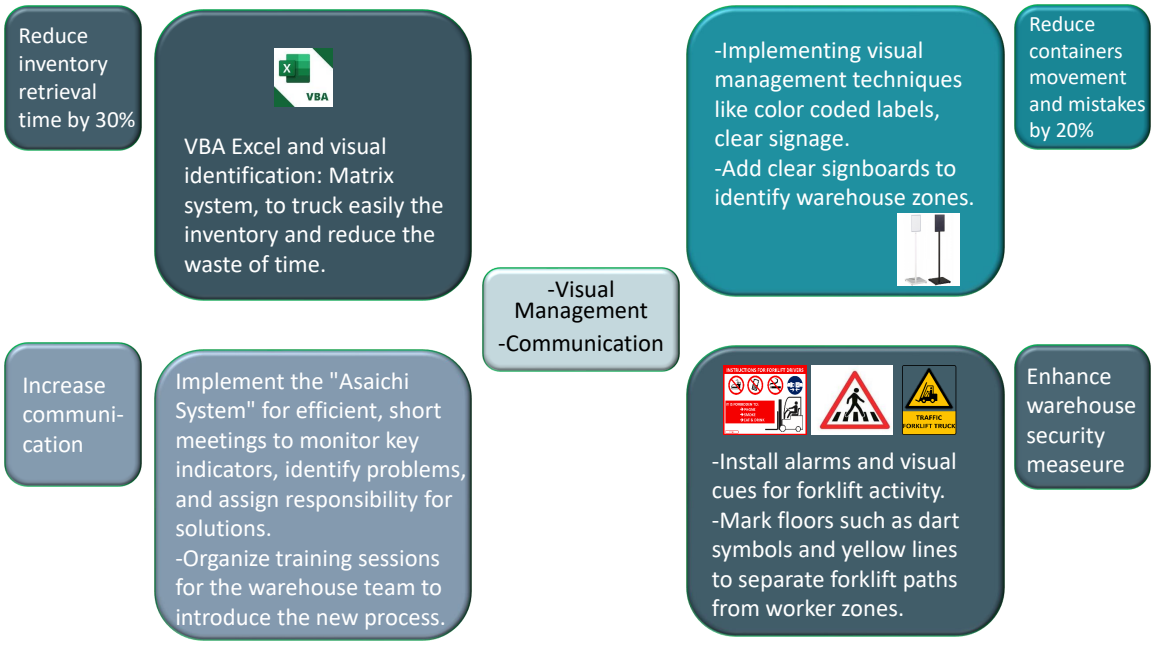



**Fig. 4.** Suggested solutions for improvement

of visual management, transport challenges and security consideration. However, it should be noted that these percentages are inclining to qualitative approaches such as staff feedback, internal benchmarking as well as observational analysis. Because there is no accurate baseline data, these values are suggestive estimates as opposed to definite quantitative outcomes. To confirm this finding, structured data collection and comparison of pre and post-implementation metrics should be made to empirically determine the size of improvement made.

## 5 Future Directions and Methodological Advancements

### 5.1 Assessing long-term effectiveness

Desirable appraisals should be conducted to measure key performance indicators (KPIs) like container movement frequency, the rate of order accuracy, and stock recovery times in a bid to maintain the suitability of the applied solutions. For instance, six months audit can show whether the changes are being maintained by comparing the baseline data and current performance indicators. Apart from performance monitoring, gaining routinely feedback from warehouse workers and overcoming new challenges along the way will make the solutions flexible to altering needs of operation. For improving the robustness of this methodology, future work will need empirical long-term follow-up studies. These may involve such systems, as quarterly or biannual evaluations to track KPI trends and detect a pattern of steady improvement. Having set in place a continuous feedback mechanism and a clearly defined review timetable, continuous improvement of performance shall be realized. This approach not only allows

for alignment of the curriculum with the warehouse's long-term strategic goals; but also empowers the operation of continuous improvement by making use of iterative learning.

### 5.2 Comparative analysis with existing methods

A review against existing warehouse management techniques shows the strengths of the method being proposed. Traditional methods include Six Sigma and 5S which carry out business using structured frameworks for optimizing processes and reducing wastes. They are just relatively expensive and a bit inflexible, especially for smaller operations. On the contrary, the proposed approach is low cost and flexible, which makes it highly relevant to Small and Medium-sized Enterprises (SMEs). Concentrating on practical tools such as Ishikawa diagrams and SMART goals, this approach will be much simpler and would not require many resources to implement. Though not as comprehensive as more complicated systems, it is just about quick enough to be implemented, and its flexibility provides an easy option towards enhancing warehouse efficiency.

## 6 Conclusion

This paper describes the importance of quality tools, lean manufacturing for improving factory operations, particularly with respect to small storage environments, with emphasis on inventory management applied to SMEs. The research is driven by the growing need to optimize warehouse efficiency, reduce costs, and manage increasing complexity in inventory systems. By examining past inefficiencies and critically evaluating current behaviors, the study presents a four-step approach to achieve these objectives.

It provides practical solutions to eliminate identified anomalies through the Ishikawa diagram and SMART goals. These solutions offer several key benefits. Economically, they reduce operational costs by minimizing excess inventory, streamlining processes, and improving space utilization. Sustainable benefits are achieved by reducing material waste and energy consumption through more efficient operations.

Maintaining long-term sustainability and continual development requires significant time and financial investment, which could lead to delays. Other critical challenges include overcoming resistance to change, ensuring data quality and consistency, and providing thorough training. Furthermore, because the research was conducted in a single organization, the methodology's application is limited by the scope of this work. To improve the results, more case studies from other businesses should be investigated, taking into account different types of storage systems and layout configurations.

Future work should focus on implementing this methodology in more complex storage systems, as well as in push production environments, to analyze how the methodology behaves in identifying and eliminating problems. Furthermore, the implementation of emerging technologies.

## Acknowledgment

This work was supported by national funds through FCT/MCTES (PIDDAC): CeDRI, UIDB/05757/2020 (DOI: 10.54499/UIDB/05757/2020) and UIDP/05757/2020

(DOI: 10.54499/UIDP/05757/2020); and SusTEC, LA/P/0007/2020 (DOI: 10.54499/LA/P/0007/2020).

## References

1. L. Wilson, "How to implement lean manufacturing," 2010.
2. Y. Monden, *Toyota production system: an integrated approach to just-in-time*. CRc Press, 2011.
3. N. Skhmot, "The 8 wastes of lean," *The lean way*, vol. 5, p. 2017, 2017.
4. C. Kuligovski, A. W. Robert, C. M. O. d. Azeredo, J. A. P. Setti, and A. M. d. Aguiar, "5s and 5w2h tools applied to research laboratories: experience from instituto carlos chagas-fiocruz/pr for cell culture practices," *Brazilian Archives of Biology and Technology*, vol. 64, p. e21200723, 2021.
5. R. Dunford, Q. Su, and E. Tamang, "The pareto principle," 2014.
6. M. Braglia, M. Frosolini, and M. Gallo, "Smed enhanced with 5-whys analysis to improve set-upreduction programs: the swan approach," *The international journal of advanced manufacturing technology*, vol. 90, pp. 1845–1855, 2017.
7. C. N. Johnson, "The benefits fo pdca," *Quality Progress*, vol. 35, no. 5, p. 120, 2002.
8. J. Singh and S. K. Gandhi, "Benefits using pdca cycle of continuous improvement in manufacturing industry-a case study," *International Journal of Management Concepts and Philosophy*, vol. 17, no. 1, pp. 83–97, 2024.

# Analysis of Liver Patients with Machine Learning

Guilherme Rodrigues[1], Gabriel A. Leite[2], Beatriz Flamia Azevedo[2], and Ana I. Pereira[2]

[1] Instituto Politécnico de Bragança, Bragança, Portugal
[2] Research Center in Digitalization and Intelligent Robotics (CeDRI), Laboratório Associado para a Sustentabilidade e Tecnologia em Regiões de Montanha (SusTEC), Instituto Politécnico de Bragança, Bragança, Portugal
a51824@alunos.ipb.pt, {gabriel.leite,beatrizflamia,apereira}@ipb.pt

**Abstract.** Liver diseases are currently a major global health problem with millions of deaths annually. This study applies a Machine Learning algorithm to identify clinical subgroups in the Indian Liver Patient Dataset. After preprocessing the data, included imputation of missing values, normalization through logarithmic transformation and feature selection through Recursive Feature Elimination with Cross-Validation, three main features (Age, Alkphosp and Sgot) were selected for clustering analysis. The $k$-means algorithm, combined with Elbow Method, outlined three distinct groups of patients. Although the groups provided clinically relevant insights, such as the association between advanced age and elevated liver enzymes, the overlap between patient groups indicated limitations in disease identification. Future work will consider implementing another clustering method, Fuzzy $c$-means, and using larger datasets.

**Keywords:** Liver Disease · Machine Learning · $k$-means · Unsupervised Learning · Indian Liver Patient Dataset

## 1 Introduction

Liver diseases are causing the death of millions of people every year, due to liver diseases such as cirrhosis, hepatitis and fatty liver disease [9]. Symptoms often only appear in advanced stages, such as fatigue and abdominal pain [1]. Additionally, liver enzyme levels (e.g., alkaline phosphatase and aspartate aminotransferase) vary from person to person, and an altered value may not always mean a serious illness, which makes detection of this type of disease a challenge [5].

Machine Learning (ML) has advanced greatly over the years, becoming a useful tool in healthcare that allows us to analyze large amounts of medical data to potentially discover patterns and predict clinical outcomes. The $k-$means algorithm has been widely used to perform unsupervised clustering that helps us identify groups of patients based on their clinical characteristics without the need for labels [7]. These methods are useful in the early stages of analysis, when the goal is to discover relationships between clinical variables and identify subgroups of patients for further research.

Several studies have used ML in the analysis of liver diseases, Spann [8] carried out a comprehensive review on the use of ML in liver diseases and transplants, highlighting the ability of algorithms such as $k$-means and Support Vector Machine (SVM) to identify clinical patterns. Ghosh and Waheed [3] conducted a study where they applied supervised algorithms, such as $k$-star, to the Indian Liver Patient

Dataset (ILPD), where they managed to achieve 100% accuracy in classifying patients with and without liver disease.

This study aims to identify patterns in clinical data of liver patients, using ML algorithms to analyze subgroups with similar characteristics, which can help in understanding the progression of the disease.

The paper is organized as follows. Section 1 provides the paper introduction. Section 2 outlines the methodology used. In Section 3, the main characteristics of the dataset are described, and Section 4 presents the obtained results. Finally, conclusions and future work are discussed in the last section.

## 2 Methodology

The methodological process involved two stages: dataset preprocessing and the application of the $k$-means algorithm. The ILPD dataset used contains 583 patient records with 11 features, as detailed in section 3. The methodology flow can be seen in the Fig. 1.

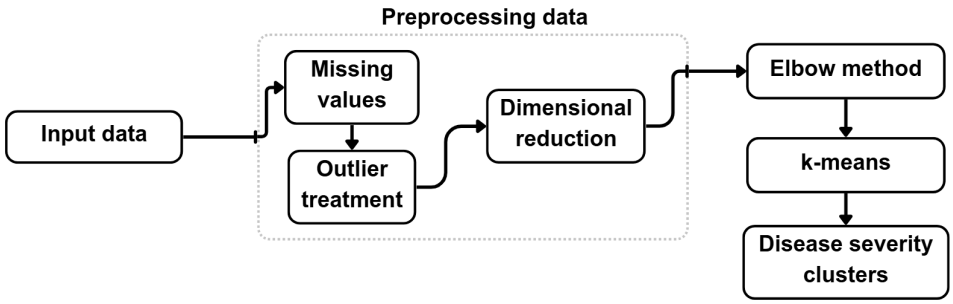



**Fig. 1.** Methodology workflow

### 2.1 Dataset preprocessing

In the initial preprocessing phase, an analysis of the presence of missing values in the dataset was performed, as the existence of missing data could later compromise the effectiveness of the $k$-means algorithm. Four null values were found in the feature Albumin and Globulin Ratio (A/G Ratio), so it was decided to perform their mean imputation.

Then, boxplots to identify outliers were made to visualize the extreme values. After visualizing the outliers, skewness graphs were also created for each feature in order to evaluate the distribution of the data. Some features were found to have significantly skewed distributions, so the natural logarithmic transformation was applied as a way to reduce this skewness and improve data normalization. However, it was decided not to remove the identified outliers, since, in clinical contexts, these may represent extreme but relevant cases, whose inclusion could enrich the segmentation carried out by the

$k$-means algorithm. Furthermore, excluding an outlier would be equivalent to excluding a patient.

Later, a dimensionality reduction was performed with the aim of selecting the most relevant features. For this purpose, a correlation matrix was constructed to evaluate the relationship between the different features of the dataset, in this phase gender was coded numerically, 0 for women and 1 for man, once the correlation matrix only works with numerical variables. Based on the correlation matrix, only one of the features with a high mutual correlation was selected, prioritizing the one with the highest correlation with the Selector variable.

To optimize the final set of features from dimensionality reduction, the Recursive Feature Elimination with Cross-Validation (RFECV) method was applied using Random Forest to train the model and identify the variables with the greatest impact on model performance.

Finally, to standardize the features before applying $k$-means, *MinMaxScaler* function was used, which transforms the data to a common scale (e.g. $[0, 1]$) and ensures that all variables have equal weight in the algorithm.

### 2.2 *K*-means clustering algorithm

$K$-means is a popular unsupervised ML algorithm used for clustering data points into distinct groups based on their similarity. It aims to separate $n$ data points into $k$ clusters, where each data point belongs to the cluster with the nearest centroid. The algorithm starts with random centroids and iteratively relocates the points and recalculates the centroids until it converges, minimizing the sum of the squares of the distances between each data point and its centroid, reaching a local minimum [4].

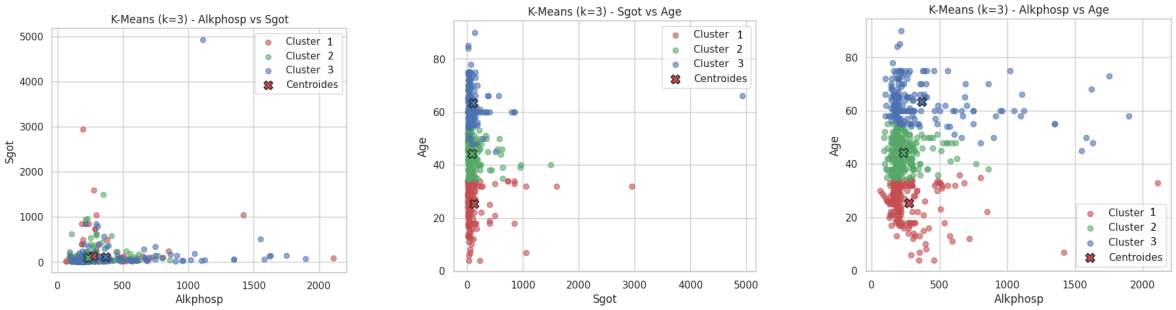
In this study, to find the optimal number of $k$ for applying the algorithm, the elbow method was used, which consists of running the clustering algorithm for different values of $k$ and measuring how far the data points are from the center of their clusters, which is called the sum of squared error. This value is then plotted on a graph, where the point at which the curve begins to flatten, forming an "elbow," signifies the optimal number of clusters. Subsequently, the performance of the clustering was evaluated using the Silhouette Score, which measures how well each data point fits within its assigned cluster compared to other clusters. This score ranges from $-1$ to 1.

## 3 Dataset

The ILPD contains 583 patient records, of which 441 are men and 142 are women. The dataset includes 11 features: age, gender, total bilirubin (TB), direct bilirubin (DB), alkaline phosphatase (Alkphosp), serum glutamic pyruvic transaminase (Sgpt), serum glutamic oxaloacetic transaminase (Sgot), total protein (TP), albumin (ALB), albumin/globulin ratio (A/G Ratio) and Selector. The Selector feature indicates the presence (1) or absence (2) of liver disease, with 416 patients and 167 healthy individuals. Patients over 89 years of age are recorded as being 90 years old.

## 4 Results and discussion

Initially, the $k$-means algorithm was applied with $k = 3$, which was calculated to be the optimal value based on the Silhouette Score and the Elbow method.




**(a)** Alkphosp vs. Sgot. **(b)** Age vs. Sgot. **(c)** Age vs. Alkphosp.

**Fig. 2.** k-means scatter plots ($k = 3$)

Although $k = 3$ was considered the most appropriated number of clusters, visual analysis of the Fig. 2 and the Table 1 indicates that the clusters did not clearly separate patients with and without liver disease, as verified by the overlap of patients with different targets values in each cluster. In an attempt to improve separation, the process was replicated, but with different $k$ values, where no significant improvement in patient separation was noted.

**Table 1.** Clusters vs Selector with ($k = 3$)

| Cluster | Selector 1 | Selector 2 |
|---|---|---|
| 1 | 112 | 56 |
| 2 | 165 | 66 |
| 3 | 139 | 45 |

In order to analyze the patients through the centroids of the clusters, which can be seen in Table 2, it is crucial to consider the range denoted as healthy for the patients for each variable considered, as suggested by [2, 6]: Serum Glutamic Oxaloacetic Transaminase (5 - 40 $\mu$/l) and Alkphos (36 - 150 $\mu$/l). The specific age range has not been defined, as there are several factors that can have an impact on this metric.

**Table 2.** Features Centroids

| Cluster | Age | Alkphosp | Sgot |
|---|---|---|---|
| 1 | 25.26 | 277.35 | 129.77 |
| 2 | 44.17 | 235.76 | 93.06 |
| 3 | 63.26 | 371.47 | 112.94 |

Based on the ranges considered healthy for Sgot and Alkphos, the clusters can be qualified as: Cluster 1, with younger patients (mean 25.26 years), Alkphosp of 277.35 $\mu$/l and Sgot of 129.77 $\mu$/l, is classified as the other cluster, since the values of Alkphosp and Sgot are significantly above the ranges considered healthy. Cluster 2, including middle-aged patients (mean 44.17 years), with the lowest values of Alkphosp (235.76 $\mu$/l) and Sgot (93.06 $\mu$/l), is considered the cluster with healthy people; although the values are above the healthy ranges, they are the least divergent between the clusters, indicating a healthier profile. Cluster 3, which includes older patients (mean 63.26 years), has the highest levels of Alkphosp (371.47 $\mu$/l) and Sgot (112.94 $\mu$/l), both significantly above healthy ranges, making the cluster with sickest people and showing serious liver problems.

## 5 Conclusion and Future Work

This study was based on the $K$-means algorithm for the analysis of liver patients in the ILPD, where it identified 3 subgroups of patients based on Age, Alkphosp and Sgot. Although the groups do not directly correspond to the identification of liver diseases, they revealed clinically relevant patterns, such as the association between advanced age and high levels of liver enzymes. The same procedure was replicated but for different values of $k$, where the results were not very different for $k = 3$.

Although the results were not very satisfactory, they are relevant to the area of liver health, as they highlight the usefulness of $K$-means, allowing the identification of subgroups that can guide clinical studies on risk factors and disease progression. However, the study found a number of limitations that might have impacted the findings, including the use of a relatively small sample, outliers, and some unbalanced data in the dataset, such as the distribution of genders (441 males vs. 141 females) and health conditions (416 patients with liver disease vs. 167 without liver disease). These limitations resulted in the skewing of the representation of clusters towards the majority classes, which could have had an impact on the clustering results. Given these characteristics, the ability of the method applied to clearly distinguish between healthy and sick patients may have been negatively affected.

For future studies, another machine learning algorithm, such as Fuzzy $c$-Means, will be applied, since this algorithm allows patients to belong to multiple clusters with different degrees of membership. Using a dataset with a larger sample could improve the quality of the clusters and the definition of the centroids, as clustering algorithms benefit from larger datasets.

## Acknowledgments

This work was supported by national funds: UID/05757 - Research Centre in Digitalization and Intelligent Robotics (CeDRI); and SusTEC, LA/P/0007/2020 (DOI: 10.54499/LA/P/0007/2020).

## References

1. Adeboye, N.O., Adesina, O.S., Afolabi, H.A., Ogunleye, T.A., Kolawole, M.K.: Advanced principal component analysis of various risk factors of hepatitis b prevalence in nigeria. Tanzania Journal of Science (2024), https://api.semanticscholar.org/CorpusID:273134693
2. Ceriotti, F., Henny, J., Queraltó, J., Ziyu, S., Özarda, Y., Chen, B., Boyd, J.C., Panteghini, M., on behalf of the IFCC Committee on Reference Intervals, (C-RIDL), D.L., on Reference Systems for Enzymes (C-RSE), C.: Common reference intervals for aspartate aminotransferase (ast), alanine aminotransferase (alt) and $\gamma$-glutamyl transferase (ggt) in serum: results from an ifcc multicenter study. Clinical Chemistry and Laboratory Medicine **48**(11), 1593–1601 (2010). https://doi.org/doi:10.1515/CCLM.2010.315, https://doi.org/10.1515/CCLM.2010.315
3. Ghosh, S.R., Waheed, S.: Analysis of classification algorithms for liver disease diagnosis. Journal of Science, Technology and Environment Informatics **05**(01), 361–3270 (2017). https://doi.org/10.18801/jstei.050117.38
4. Ikotun, A.M., Ezugwu, A.E., Abualigah, L.: K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. Information Sciences **622**, 178–210 (2023). https://doi.org/https://doi.org/10.1016/j.ins.2022.11.139, https://www.sciencedirect.com/science/article/pii/S0020025522014633
5. Johnston, D.E.: Special considerations in interpreting liver function tests. American Family Physician **59**(8), 2223–2230 (1999)
6. Sharma, U., Pal, D., Prasad, R.: Alkaline phosphatase: an overview. Indian J Clin Biochem **29**(3), 269–278 (Nov 2013)
7. Shukla, S.: A review on k-means data clustering approach. International Journal of Information and Computation Technology **4**(17), 1847–1860 (2014), https://api.semanticscholar.org/CorpusID:17937058
8. Spann, A., Yasodhara, A., Kang, J., Watt, K., Wang, B., Goldenberg, A., Bhat, M.: Applying machine learning in liver disease and transplantation: A comprehensive review. Hepatology **71** (2020). https://doi.org/10.1002/hep.31103
9. World Health Organization: Who sounds alarm on viral hepatitis infections claiming 3,500 lives each day (2024), https://www.who.int/news/item/09-04-2024-who-sounds-alarm-on-viral-hepatitis-infections-claiming-3500-lives-each-day

# Implementation of an Asset Administration Shell Type 3 in an Automotive Assembly Line

José Costa[1], Lucas Sakurada[1], and Paulo Leitao[1]

Research Center in Digitalization and Intelligent Robotics (CeDRI), Laboratório Associado para a Sustentabilidade e Tecnologia em Regiões de Montanha (SusTEC), Instituto Politécnico de Bragança, Campus de Santa Apolónia, 5300-253 Bragança, Portugal
{jose-costa, lsakurada, pleitao}@ipb.pt

**Abstract.** The Asset Administration Shell (AAS) is a standardised digital representation that bridges the physical and digital worlds and is designed to enable data exchange and interoperability in Cyber-Physical Production Systems (CPPSs). Despite their benefits, the AAS faces modelling, application, and integration challenges, particularly in implementing the proactive AAS (Type 3). This preliminary work describes the implementation of an AAS Type 3 in an automotive assembly line, considering the Multi-Agent Systems (MASs) principles to transform an AAS Type 2 into an AAS Type 3. Preliminary results demonstrate that, using MAS principles in a Type 3 AAS, it is possible to reduce measurement time and, through agent intelligence, expand the capacity of the AAS.

**Keywords:** Asset Administration Shells · Cyber-Physical Production Systems · Industry 4.0 · Multi-Agent System

## 1 Introduction

The Asset Administration Shell (AAS) emerged as a significant concept in the Industry 4.0 (I4.0) context, as defined by the Reference Architectural Model Industrie 4.0 (RAMI4.0) [1, 3, 13] to create an interface between the physical and virtual worlds [2], acting as a single source of truth within the assessment process and creating interoperability between assets [14].

The AAS concept is a standardised framework designed to represent, manage, and facilitate the interchange of physical asset-related data in an interoperable and machine-readable manner [7]. The AAS can be classified into three different types according to their level of interaction capability and degree of autonomy to make decisions [8], namely passive AAS (Type 1), which is a file-based data structure that standardises asset information; reactive AAS (Type 2) acts as an Application Programming Interface (API) that exchanges information via HTTP/REST protocol; and, finally, proactive AAS (Type 3) acts as a decision-making enables communication and collaboration with other AASs (e.g., through negotiation).

Although several AAS Type 1 and Type 2 implementations exist and have been successfully demonstrated in industrial environments, namely [15–17], developing AAS Type 3 remains challenging [9,18]. The AAS Type 3 requires further research to address autonomous decision-making, context awareness, and dynamic coordination between heterogeneous assets [12].

With this in mind, this preliminary work describes the implementation of an AAS Type 3 for an automotive assembly line, using the Multi-Agent System (MAS) principles to solve interoperability issues and enhance the pro-activeness of traditional AAS.

This article is organised into five sections. After the introductory section, Section 2 provides the agent-based AAS architecture concept. Section 3 describes the automotive assembly line use case. Section 4 explains how the AAS Type 1, AAS Type 2 and the agent-based AAS are modelled and provide the tests, the results and their discussion. Finally, section 5 concludes the paper and presents the potential future work.

## 2 Agent-based Asset Administration Shell Architecture

As aforementioned, AAS Type 3 represents an advanced and extended form of AAS but faces a gap in terms of interaction between AAS. In fact, the definition of AAS Type 3 recalls MAS, where each AAS functions as an autonomous, intelligent, and cooperative entity capable of interacting with other AASs to exchange information and make decisions. Currently, AAS Type 3 is still in the early stages of development and lacks specific guidelines and specifications for its implementation. In this context, this section addresses this gap by introducing a conceptual architecture that integrates MAS and AAS to realise AAS Type 3 which forms an agent-based AAS architecture [11], Fig. 1 illustrates the general concept of the proposed agent-based AAS architecture.

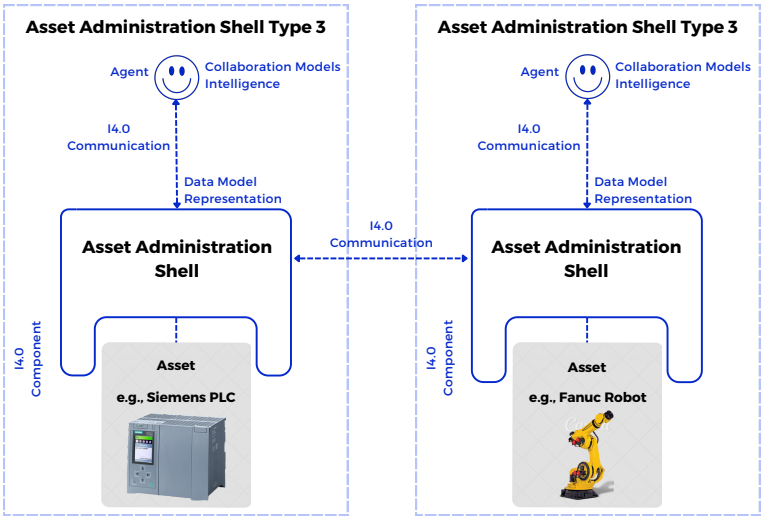



**Fig. 1.** General concept of agent-based AAS architecture.

The agent-based AAS architecture aims to enhance digitisation by embedding distributed intelligence and autonomy of MASs and collaborative functions supporting interoperability, flexibility, adaptability, and interoperability requirements of I4.0 [11]. MAS technology is used to realise intelligent AAS, mapping their inherent characteristics into AAS functionalities and extending them with intelligence and data analytics capabilities [12].

While AAS developments have focused on information management, there is a growing need to incorporate intelligence and collaboration to develop I4.0 compliant solutions. Integrating MAS technology is a step towards addressing this gap, although

such solutions are still in the early stages of maturity [10, 11]. In conclusion, the agent-based AAS architecture is a promising approach to advancing the digitisation and intelligence of industrial assets, enhancing collaboration and adaptability in manufacturing systems. However, further development and validation are needed to realise its full potential in I4.0 environments.

## 3 Description of the Case Study

As shown in Fig. 2, the case study is related to an automotive assembly line, particularly the final assembly. The measurements of the flush and gap of the automobile are carried out using an "intelligent" device developed by U.Sense during the openZDM project, named G3F, which allows for making the measurements using a laser and storing them automatically.

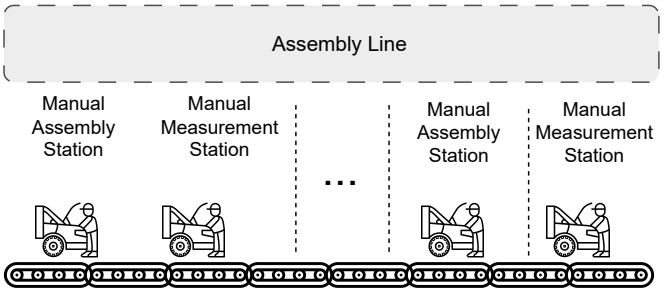



**Fig. 2.** Automotive assembly line use case.

However, the proper laser measurements require knowing the car's colour to adjust the laser. Additionally, the device needs to know the defined limits for each measurement to evaluate if the gaps and flushes are within the tolerance and rejection limits. In this context, it is necessary that the G3F sensor and the cars, modelled as AAS, should exhibit proactive behaviours (Type 3) and not single reactive ones.

## 4 Experimental Implementation

This section describes the experimental implementation of agent-based AAS, demonstrating how AAS Types can be applied to an automotive assembly line.

### 4.1 Implementation of AAS Type 1 and Type 2

The AAS Type 1 acts as a static file where the asset's information is stored in a JSON or XML file and can be exchanged digitally across the network [10]. In this preliminary work, to implement the G3F sensor and the automotive as AAS Type 1, the software Eclipse AASX Package can be used, which consists of viewing, creating, editing and hosting an I4.0 component [6, 11].

An AAS is modular and comprises several `submodels`, which refer to models that are technically separate from each other and included in the AAS [6]; it can consist of data such as identification, technical data, operational data and capabilities. The G3F sensor has a `submodel` called `OperationalData`, which includes the `car_id`, `colour`, and `MeasurementData`. Similarly, the Car includes the `submodel` `ProductSpecification`, responsible for containing specific operational data such as `OperationalDataG3F`.

To initialise them as AAS Type 2, the Eclipse AASX Server can be used, which provides the necessary infrastructure to access real-time data, allowing communication via HTTP/REST and initiated by the client.

### 4.2 Implementation of Multi-Agent System Infrastructure

When implementing an AAS Type 3, each AAS must demonstrate collaborative and intelligent capabilities. To support these capabilities, software agents are introduced to manage interaction, coordination, and negotiation between AASs. This agent-based approach enables each AAS to behave proactively and to exchange information with other AASs autonomously through their respective agents. Fig. 3 illustrates how this architecture is applied in the automotive assembly line use case, highlighting the role of agents in facilitating distributed decision-making and cooperation.

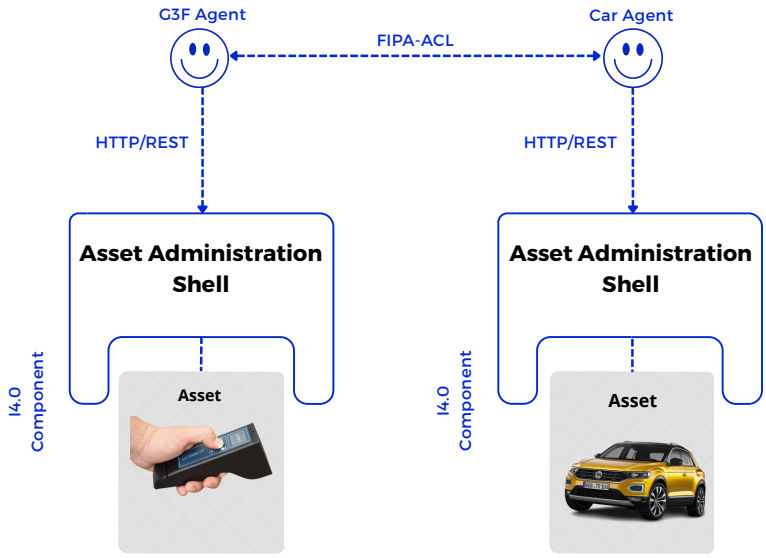



**Fig. 3.** Agent-based AAS architecture for automotive assembly line use case.

The agents were developed in Java Agent DEvelopment Framework (JADE) (jade.tilab.com), implementing an efficient agent platform and supporting the development of MAS [4]. JADE is known for its compliance with the Foundation for Intelligent Physical Agents (FIPA) (fipa.org) standards [5]; having this in mind, the agent-based AAS architecture is composed of two principal agents:

– G3F Agent: is responsible for managing the `car_id` readings of each Car and the measurements of each point name.
– Car Agent: is responsible for managing the catalogue of cars by using the available resources (following a plan containing each `car_id`, `colour`, and `MeasurementData`).

The G3F Agent's behaviour is to collect real-time data from the G3F sensor. It is responsible for identifying the car on the assembly line via its `car_id`, and requesting the necessary data from the Car Agent. On the other hand, the Car Agent represents the car and is responsible for managing and making available its technical data and responding to requests from the G3F Agent.

### 4.3 Interaction Protocol

The interaction between agents in the JADE platform is based on the Agent Communication Language (ACL) specified by the FIPA standard [5]. Following Fig. 4, the sequence starts when the G3F sensor publishes the `car_id` in MQTT and the G3F Agent starts the interaction by sending a `REQUEST` message to the predefined agent that implements the Yellow Pages [4], the Directory Facilitator (DF), asks which agents provide the `car_id`. The DF returns the Agent Identifier (AID) in an `INFORM` message.

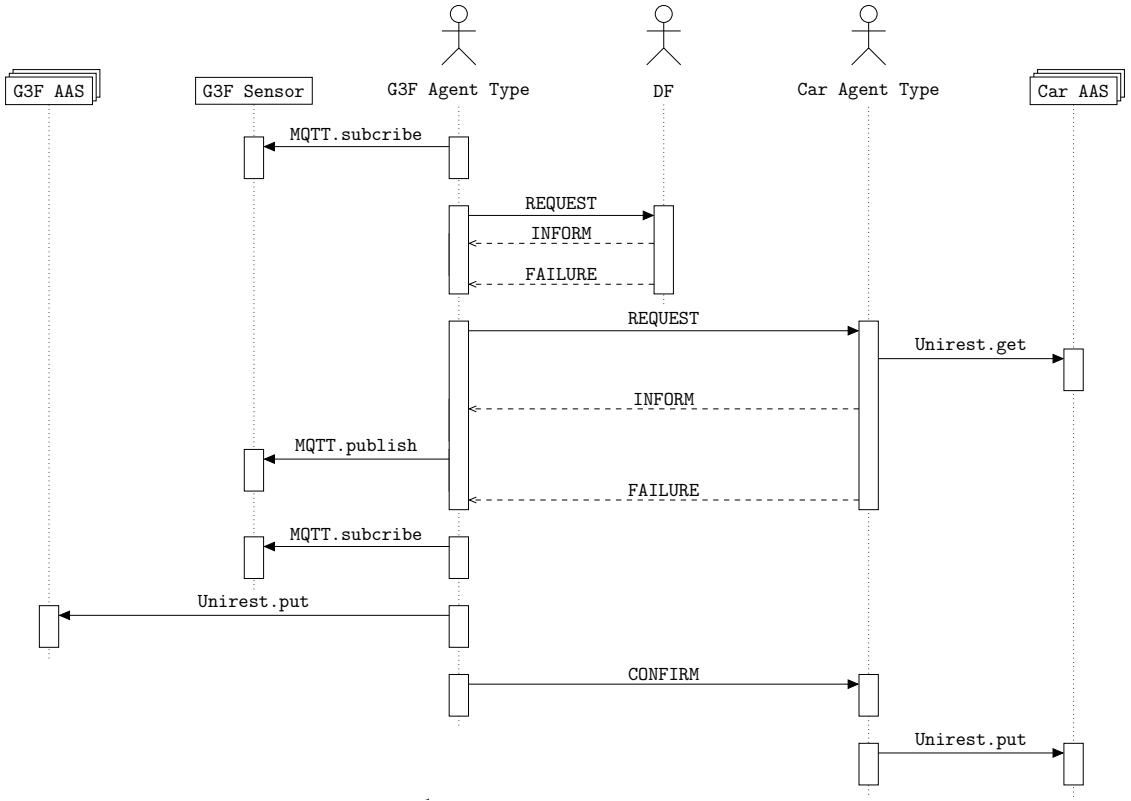



**Fig. 4.** Interaction protocols between sensor, agents and AAS.

After that, when the G3F Agent receives the information, the agent asks the AID provider for the `colour`, `limits_points` and `point's_names`; this data is contained in the Car AAS, so the Car Agent needs to connect to its AAS via HTTP/REST protocol, execute the `get` command and then inform the G3F Agent. G3F Agent then publishes these values in the MQTT, and the human operator with the G3F sensor takes the gap and flush measurements. When all the `point's_names` have been measured, and the G3F sensor publishes them in the MQTT, and the G3F Agent executes the `put` command in its own G3F AAS and confirms the Car Agent of the measurements to do the same.

### 4.4 Analysis of Results

This preliminary work was experimentally implemented and features modularity and scalability, as it demonstrates the integration of MAS principles into AAS, allowing the transition from AAS Type 2 to AAS Type 3. The agents could autonomously exchange and act upon information, enhancing the flexibility and responsiveness of the assembly line operations. The architecture supports interoperability and real-time decision-making, key requirements for I4.0 environments.

The response time of the interactions between the G3F Agent and the Car Agent is shown in Fig. 5, which shows the results for 40 experiments with the average time being 0.0565 seconds and a pattern deviation of 0.0671 seconds. In this case, the pattern deviation is slightly higher than the average because there are outliers; for example, the initial value of 0.446 seconds is considerably higher than the others, which increases it significantly. The increase in response time in trials 27 to 40 is because the DF is being used as a resource for the G3F Agent, and there was a continuous request, so the DF took longer to process the data.

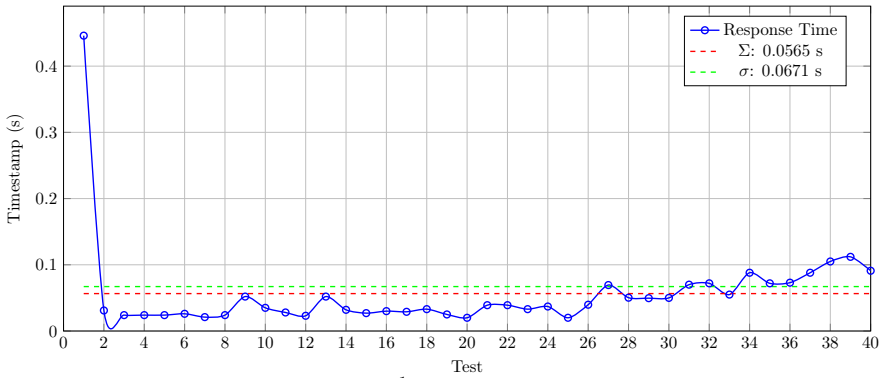



**Fig. 5.** Response time between agents.

## 5 Conclusion

This preliminary work describes implementing an agent-based AAS approach to improve interoperability and autonomy in an automotive assembly line by leveraging

MAS, the proposed architecture transitioned from AAS Type 2 to Type 3, allowing the real-time adaptability.

This preliminary work showed that by using an agent-based AAS architecture, however it is more complex than a function-based architecture in terms of modelling and development, by integrating agents into AASs, it is possible to implement proactive behaviour with the ability to locate, negotiate and exchange information, without relying on a centralised architecture. With this in mind, interoperability and modularity have been made possible by integrating agents, ensuring effective communication across the AASs.

Future work will focus on improving communication performance and expanding the use to more intricate industrial scenarios.

## Acknowledgment

This work was partially supported by the HORIZON-CL4-2021-TWIN-TRANSITION-01 openZDM project under Grant Agreement No. 101058673 and was also supported by national funds through FCT/MCTES (PIDDAC): CeDRI, UIDB/05757/2020 (DOI: 10.54499/UIDB/05757/2020) and UIDP/05757/2020 (DOI: 10.54499/UIDP/05757/2020); and SusTEC, LA/P/0007/2020 (DOI: 10.54499/LA/P/0007/2020).

## References

1. Arm, J., Kaczmarczyk, V., Benesl, T., Marcon, P., Jirgl, M., Bradac, Z.: Asset Administration Shell as the Key Enabler of the Industry 4.0 Phenomenon. IFAC-PapersOnLine **58**(9), 275–280 (2024), 18th IFAC Conference on Programmable Devices and Embedded Systems PDES 2024
2. Arm, J., Benesl, T., Marcon, P., Bradac, Z., Schröder, T., Belyaev, A., Werner, T., Braun, V., Kamensky, P., Zezulka, F., et al.: Automated Design and Integration of Asset Administration Shells in Components of Industry 4.0. Sensors **21**(6), 1 – 20 (2021)
3. DIN: Reference Architecture Model Industrie 4.0 (RAMI4.0) English translation of DIN SPEC 91345:2016-04 (2016)
4. Fabio Luigi Bellifemine, Giovanni Caire, D.G.: Developing Multi-Agent Systems with JADE. John Wiley & Sons, Ltd (2007)
5. FIPA: FIPA ACL Message Structure Specification
6. Industrial Digital Twin Association: Specification of the Asset Administration Shell - Part 1: Metamodel (version 3.0.1) (2024)
7. Kosse, S., Betker, V., Hagedorn, P., König, M., Schmidt, T.: A Semantic Digital Twin for the Dynamic Scheduling of Industry 4.0-based Production of Precast Concrete Elements. Advanced Engineering Informatics **62**, 102677 (2024)
8. Plattform Industrie 4.0: Verwaltungsschale in der Praxis (2020)
9. Sakurada, L., De La Prieta, F., Leitao, P.: A Methodology for Integrating Asset Administration Shells and Multi-Agent Systems. In: Proceedings of the 2023 IEEE 32nd International Symposium on Industrial Electronics (ISIE). pp. 1–6. IEEE (2023)
10. Sakurada, L., Leitao, P., De La Prieta, F.: Engineering a Multi-Agent Systems Approach for Realizing Collaborative Asset Administration Shells. In: Proceedings of the 2022 IEEE International Conference on Industrial Technology (ICIT). pp. 1–6. IEEE (2022)
11. Sakurada, L., Leitao, P., la Prieta, F.D.: Agent-Based Asset Administration Shell Approach for Digitizing Industrial Assets. IFAC-PapersOnLine **55**(2), 193–198 (2022), 14th IFAC Workshop on Intelligent Manufacturing Systems IMS 2022

12. Sakurada, L., Leitão, P., de la Prieta, F., Corchado, J.M.: Multi-Agent Systems to Realize Intelligent Asset Administration Shells. In: Proceedings of the III Workshop on disruptive information and communication technologies for innovation and digital transformation. pp. 43–58. Ediciones Universidad de Salamanca (2021)
13. Sakurada, L., De la Prieta, F., Leitao, P.: Digitization of Industrial Environments through an Industry 4.0 Compliant Approach. In: IECON 2023-49th Annual Conference of the IEEE Industrial Electronics Society. pp. 1–6. IEEE (2023)
14. Shi, D., Liedl, P., Bauernhansl, T.: Interoperable Information Modelling Leveraging Asset Administration Shell and Large Language Model for Quality Control toward Zero Defect Manufacturing. Journal of Manufacturing Systems **77**, 678 – 696 (2024)
15. Ye, X., Hong, S.H.: Toward Industry 4.0 Components: Insights Into and Implementation of Asset Administration Shells. IEEE Industrial Electronics Magazine **13**(1), 13–25 (2019)
16. Ye, X., Hong, S.H., Song, W.S., Kim, Y.C., Zhang, X.: An Industry 4.0 Asset Administration Shell-Enabled Digital Solution for Robot-Based Manufacturing Systems. IEEE Access **9**, 154448–154459 (2021). https://doi.org/10.1109/ACCESS.2021.3128580
17. Ye, X., Jiang, J., Lee, C., Kim, N., Yu, M., Hong, S.H.: Toward the Plug-and-Produce Capability for Industry 4.0: An Asset Administration Shell Approach. IEEE Industrial Electronics Magazine **14**(4), 146–157 (2020). https://doi.org/10.1109/MIE.2020.3010492
18. Zhang, J., Ellwein, C., Heithoff, M., Michael, J., Wortmann, A.: Digital twin and the asset administration shell. Software and Systems Modeling pp. 1–23 (2025). https://doi.org/10.1007/s10270-024-01255-0

# Security Threat Modeling for Identifying Vulnerabilities in a Hate Speech Detection System Based on NLP

Ruth Mendonça[1,3], Gustavo Funchal[2], Frederico Barbosa Muniz[3], and Tiago Pedrosa[2]

[1] Instituto Politécnico de Bragança, Bragança, Portugal
[2] Research Centre in Digitalization and Intelligent Robotics (CeDRI), Laboratório Associado para a Sustentabilidade e Tecnologia em Regiões de Montanha (SusTEC), Instituto Politécnico de Bragança (IPB), Bragança, Portugal
m321979@alunos.ipb.pt, {gustavofunchal, pedrosa}@ipb.pt
[3] Faculdade de Tecnologia do Estado de São Paulo (Fatec Registro)
frederico.muniz@fatec.sp.gov.br

**Abstract.** The growing increase in hate speech in digital environments, especially when accessed by students, constitutes a significant threat to inclusive education. Natural Language Processing (NLP) systems have been used as tools for detecting and mitigating this type of content, but these systems are also subject to security risks, such as manipulation, adversarial attacks or internal flaws. This paper proposes the application of the STRIDE threat modeling framework to the flowchart of the process of a hate speech detection system designed for school environments. The aim is to identify security weaknesses in the components of this system, from data ingestion and model inference to user interaction, as well as to propose mitigation strategies to increase the system's reliability. By aligning threat modeling practices with the specific context of offensive language detection, this study contributes to the development of more secure and effective content auditing systems.

**Keywords:** Threat modeling · Hate speech detection · STRIDE · NLP · Security in AI systems

## 1 Introduction

Hate speech on the Internet has become increasingly prevalent, in line with the growth in the use of social networks, which currently reach around 60% of the world's population [2]. Although these digital spaces encourage freedom of expression and the sharing of ideas, the anonymity that characterizes them often facilitates offensive behaviour, resulting in humiliation, harassment and discriminatory practices. This online toxicity, often referred to as cyberbullying, represents a constant threat to safety in the digital environment and to the right to participate in virtual spaces.

The global scale of the problem becomes evident when considering reports such as [1], where the authors state that 80% of Internet users in the European Union have witnessed hate speech online, and 40% have personally felt attacked. In [2], data from South Korea show that 64% of adults and 68.3% of young people have been exposed to this type of content, especially targeting women, immigrants, and people with disabilities. In [5], a study in Brazil found that, in an analysis of 145 news stories published in a single day, 90% contained at least one hateful comment.

This panorama becomes even more delicate when transposed to the school context, which has also been impacted by these harmful dynamics [3]. Unrestricted access to

the Internet in educational institutions exposes children and adolescents to discriminatory content, compromising both their emotional well-being and the quality of their learning. The presence of hate speech in digital school environments directly threatens the construction of a welcoming and inclusive space for all students. Faced with this scenario, the Information Technology (IT) area has assumed a strategic role, especially through advances in Artificial Intelligence (AI). In particular, progress in Natural Language Processing (NLP) has enabled the development of automated systems capable of detecting, classifying, and blocking harmful content on the web. These systems are able to analyze large volumes of data in real time, helping to reduce the spread of hate speech and promote safer virtual environments, including in the school context. In project, where STRIDE (acronym for Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, and Elevation of Privilege) was applied, NLP is used to analyze the textual content of websites accessed by students, identifying patterns linked to hate speech. The AI system is based on the pre-trained model Bidirectional Encoder Representations from Transformers (BERT), developed by Google, known for its ability to understand word context and linguistic nuances. Training was conducted using the Jigsaw Unintended Bias dataset, which includes millions of labeled online comments across various offensive categories. The data underwent preprocessing steps such as text cleaning and tokenization to align with BERT requirements. The system classifies toxicity levels into categories like toxic, obscene, and threat. If the accuracy of the classification of the categories is equal to or greater than 50%, the content is added to a blocklist on the server to prevent future access.

However, the adoption of AI-based technologies also requires attention to security aspects. Hate speech detection tools can be the target of attacks that affect their performance, either by exploiting flaws in their process or by subtle manipulations in the input texts, which can compromise their effectiveness. To deal with these threats, the application of threat modeling techniques is essential, making it possible to identify possible vulnerabilities and propose mitigation measures from the initial stages of the project.

One notable approach for addressing security threats is the STRIDE model, proposed in [7]. This model is especially useful in different contexts, including systems that use AI. Its application in NLP solutions makes it possible to anticipate risks and strengthen the system's defenses against various types of attack, promoting greater reliability and resilience.

Having this in mind, this paper proposes the application of STRIDE threat modeling to the flowchart of the process of a system being developed to detect hate speech in school environments. The aim is to identify critical vulnerabilities and suggest improvements to ensure both the security and effectiveness of the solution in the educational context.

The remaining paper is organized as follows: Section 2 presents related work, with an emphasis on work that uses STRIDE model to analyze and model threats in AI-based systems. Section 3 describes the flowchart of the process system, highlighting its main components. Section 4 details the application of threat modeling based on the STRIDE

model in the flowchart of the process presented. The 5 section discusses the results obtained, concludes the article and proposes directions for future work.

## 2 Related Work

The security of NLP systems has become an increasingly important topic, particularly as AI-based technologies are adopted more widely. Breaches in these systems, either through input manipulation or by exploiting model weaknesses, can have serious consequences. As a result, it is critical to incorporate threat modeling and risk analysis into the development of NLP applications to maintain their integrity and reliability. Several studies have proposed different strategies to evaluate and address security risks in NLP, with the STRIDE framework emerging as a commonly used approach.

In [4], the authors adapted the traditional STRIDE model to better address unique vulnerabilities in AI systems, focusing especially on input data and threats to model integrity. Their work demonstrated how STRIDE can be applied to detect flaws related to input manipulation techniques, including data poisoning and adversarial attacks. This adaptation expanded STRIDE's relevance to AI systems, highlighting its effectiveness in mapping vulnerabilities and proposing specific mitigation strategies.

Similarly, in [8], STRIDE was applied to systems built on Large Language Models (LLMs), investigating threats such as prompt injection and data poisoning. The study emphasized the usefulness of STRIDE in uncovering security gaps that arise from the context in which NLP models operate and stressed the importance of protecting both user interactions and data quality.

Beyond model-specific analyses, in [6], the authors proposed a taxonomy of attacks targeting Machine Learning systems at various stages, calling attention to the need for a comprehensive understanding of system architecture. Together, these studies suggest that threat modeling efforts must consider all parts of an AI-based system, from data input to infrastructure, to be truly effective.

While existing research has shown the value of STRIDE in AI contexts, many studies have focused primarily on isolated aspects, such as input vulnerabilities or model behavior. In this way, the proposed work aims to extend the application of STRIDE to cover the entire architecture of an offensive speech detection system, from data ingestion to user-facing outputs. This broader analysis aims to identify vulnerabilities not only within individual components, but also those emerging from interactions between them, an especially important consideration in educational environments, where safeguarding user engagement is critical.

## 3 Flowchart of the process and Threat Modeling

The flowchart of the process illustrated in Fig. 1 describes the hate speech detection system, whose structure will be evaluated using the threat modeling proposed in this study. The system integrates interconnected modules that automatically monitor and block potentially offensive content.

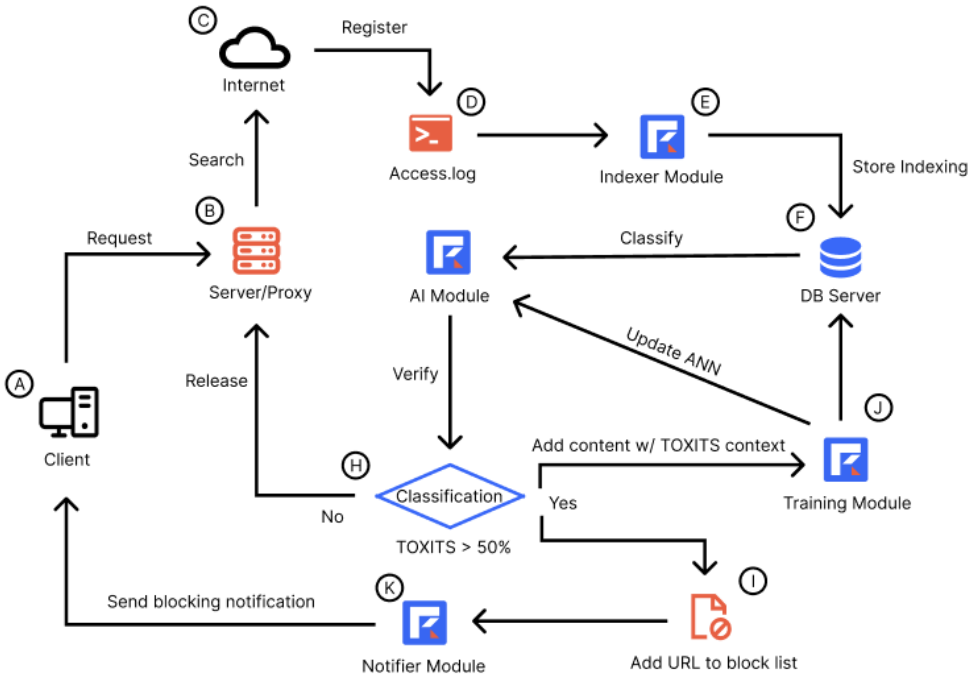



**Fig. 1.** flowchart of the process for hate speech detection and response.

The flow begins with the client's request for access (a), mediated by a proxy server (b). Requests to the Internet (c) are recorded in logs (d), which provide data to the indexer module (e). The indexer then extracts relevant information and stores it in the database (f). Subsequently, the AI module classifies the accessed content based on toxicity parameters (g). If the value identified exceeds the threshold defined as greater than 50% (h), the system automatically blocks access, sends a notification to the user via the notifier module (k), and adds the URL to the block list (i). At the same time, this content is forwarded to the training module (j), which updates the ANN model to improve its detection capacity over time. In this way, the approach ensures an agile and adaptive response to offensive content.

## 4 Threat Modeling Based on the STRIDE Model

This section uses the STRIDE methodology to assess security risks in the system architecture, identifying possible weaknesses. The framework categorizes threats into six main groups: spoofing, tampering, repudiation, information disclosure, denial of service, and elevation of privilege. The assessment focused on the vulnerabilities with the greatest potential to affect the continuous operation and reliability of the system, also considering attack scenarios and plausible malicious agent profiles.

### 4.1 Spoofing

Given that the system operates passively in the background, monitoring network traffic without requiring direct user authentication, the identification of devices in the system occurs via MAC address. Although this identification occurs at the hardware level (and not the user's), it establishes a verifiable relationship between browsing activities and authorized devices on the network.

However, one possible attack scenario involves a tech-savvy student trying to manipulate their IP address or device information to circumvent restrictions associated with their identity on the network. Even on a restricted network, the attempt to impersonate another device or mask one's digital identity is still a threat. The motivation would be to gain access to blocked content or hide their online activity, especially in institutions with less robust network infrastructure.

### 4.2 Tampering

In terms of integrity violation, logs and data processed by the indexing module (represented by (d) and (e) in the flowchart of the process presented in Fig.1, respectively) are vulnerable to manipulation if they are not protected by encryption or integrity checks. An attacker with access to the system could tamper with access logs or indexed data, distorting the history and impairing content clas- sification. Such a scenario would compromise the reliability of the system and, consequently, the effectiveness of the AI model. An attack of this type can be carried out by an internal agent, such as a technician with privileged access or a student who has compromised a terminal. The motivation can range from concealing misbehavior to sabotaging the system, making it difficult for the school administration to monitor.

### 4.3 Repudiation

The lack of robust auditing mechanisms, such as immutable logs or digital signatures, would allow users to contest recorded actions, such as attempts to access blocked URLs. This weakness would hinder accountability and incident management, reducing the administrative effectiveness of the system.

In a typical scenario, a student could contest the record of access to blocked content, claiming that browsing was accidental or fraudulent. The absence of verifiable records would allow such claims to gain traction, with the aim of avoiding disciplinary punishment and challenging the school's ability to hold users accountable based on the system's data.

### 4.4 Information Disclosure

The system processes sensitive data such as browsing histories, toxicity ratings and training sets. Modules such as the indexer and the database (represented by (e) and (f) in the flowchart of the process presented in Fig.1, respectively) can become leak vectors if exposed due to configuration failures, unprotected endpoints, or lack of encryption. Although the operation is in the background, operational or infrastructure failures can compromise users' privacy.

A possible attack scenario involves an external attacker exploiting a known vulnerability or an internal user with unauthorized access. Motivation can include collecting personal data and exploiting compromising information. In institutions with less administrative control, even accidental leaks through unmonitored unauthorized access pose an imminent risk.

### 4.5 Denial of Service

Although the system works transparently, denial of service attacks are possible, especially if malicious users overload the system with requests containing content classified as toxic against the proxy or the classifier (represented by (b) and (g) in the flowchart of the process presented in Fig.1, respectively). In controlled environments, the probability is low, but a successful attack could jeopardize the availability and stability of the system. Malicious users could automate or orchestrate access to pages known to contain blockable content, testing the limits of the infrastructure. Motivation could include trying to destabilize the system.

### 4.6 Elevation of Privilege

The system's training module (represented by (j) in the flowchart of the process presented in Fig.1), responsible for updating the AI model with new data, represents a critical point in the architecture. If there is no adequate isolation between this module and the database (represented by (f) and (e) in Fig.1), or if the input data is not validated, the system becomes vulnerable to data poisoning attacks. In this context, a malicious agent can insert manipulated examples into the database, such as offensive phrases labeled as harmless. This could compromise the AI's learning process, causing the system to misclassify content in the future. This attack could be carried out by a careless administrator, an employee with privileged access or even an external attacker exploiting loopholes in update scripts.

### 4.7 Summary

Considering the results obtained from the threat analysis carried out, it was possible to summarize the main threats for each category, as well as the severity of the threats, which are listed in Table 1.

## 5 Analysis of Results and Conclusions

STRIDE-based analysis revealed significant vulnerabilities in the hate speech detection system, particularly involving authentication, data integrity, and system reliability. MAC address identification, for example, does not require strong authentication and can be easily bypassed using widely available spoofing tools. The system's use of sensitive information—such as browsing histories and toxicity ratings—reinforces the need for safeguards such as encryption, immutable logs, and granular access control.

**Table 1.** Consolidated STRIDE-based threat modeling and analysis

| **STRIDE Category** | **Description** | **Modus Operandi** | **Severity** |
|---|---|---|---|
| Spoofing (Impersonation) | Unauthorized users impersonate teachers or administrators to bypass access controls. | An attacker might use `macchanger -r eth0` in Linux to spoof a device's MAC address and access restricted content. STRIDE recommends identity verification mechanisms. | High |
| Tampering | Injection or manipulation of input data to deceive the classifier (e.g., adversarial text). | Attackers may access the NoSQL database to alter toxicity labels or introduce adversarial examples. STRIDE led to implementation of access control and audit policies. | High |
| Repudiation | Lack of proper logging allows denial of content submission or events. | Without immutable logs, attackers may deny submitting harmful content. STRIDE emphasizes protected, verifiable logs. | Medium |
| Information Disclosure | Exposure of confidential classification results or user data. | Misconfigured networks can be exploited (e.g., via Wireshark) to leak user data. STRIDE recommends encryption and secure communication. | High |
| Denial of Service | System overload via repeated or adversarial requests. | Attackers flood the classifier with toxic inputs or long texts, degrading performance. STRIDE modeling highlighted this vulnerability. | Medium |
| Elevation of Privilege | Gaining admin access or influencing model behavior. | Malicious input via open APIs (e.g., neutral-labeled toxic phrases) may corrupt the model. STRIDE advises strict input validation and API protection. | High |

DoS attacks, while less expected on internal networks, were shown to be possible. STRIDE highlighted that attackers could intentionally generate repeated requests to URLs known to host toxic content, overwhelming the rating module and impacting service availability. Without structured threat modeling, such a scenario could have gone undetected.

One of the most serious issues identified was the risk of privilege escalation through data poisoning during AI training. A compromised model could lose accuracy, become biased, or even be manipulated — undermining the system's core function. The analysis also showed that insider threats, even in a controlled school environment, can be just as relevant as external threats.

The adoption of STRIDE was essential not only to detect technical flaws, but also to understand how system components, threat categories, and attacker profiles interact. This understanding is crucial to informing protection strategies and ensuring system integrity, and reinforces the importance of threat modeling in NLP-based solutions.

For example, **denial of service** was not initially regarded as a relevant threat, as the system operates within a controlled school network. However, STRIDE revealed that an attacker could intentionally generate repeated requests to URLs known to contain toxic content. This behavior could overwhelm the content classification modules and compromise service availability. Without the methodology, such a

scenario might not have been considered. In the future, the work will continue with the implementation of mitigation solutions for the vulnerabilities identified, as well as improving threat modeling to strengthen the system's protection and exploring more advanced approaches to ensure its security.

## References


1. Castaño-Pulgarín, S., Suárez-Betancur, N., Tilano Vega, L., Herrera López, H.: Internet, social media and online hate speech: systematic review. Aggression and Violent Behavior **58**, 101608 (2021). https://doi.org/10.1016/j.avb.2021.101608
2. Lee, S.H., Lee, J.I., Jung, G.J., Cho, H.I., Han, S.H., Hong, S.S.: Report on hate speech (2019), `https://www.humanrights.go.kr/download/BASIC_ATTACH?storageNo=1068506`
3. Livingstone, S., Smith, P.K.: Annual research review: Harms experienced by child users of online and mobile technologies: the nature, prevalence and management of sexual and aggressive risks in the digital age. Journal of Child Psychology and Psychiatry **55**(6), 635–654 (2014). https://doi.org/10.1111/jcpp.12197
4. Mauri, A., Scandariato, R., Martinelli, F., Matteucci, I.: Modeling threats to ai-ml systems using stride. Sensors **22**(17), 6662 (2022). https://doi.org/10.3390/s22176662
5. Pelle, R., Moreira, V.: Offensive comments in the brazilian web: a dataset and baseline results (2017), `https://sol.sbc.org.br/index.php/brasnam/article/view/3260`
6. Pitropakis, N., Panaousis, E., Giannaka, E., Anastasiadis, E., Loukas, G.: A taxonomy and survey of attacks against machine learning. Computer Science Review **34**, 100199 (2019). https://doi.org/10.1016/j.cosrev.2019.100199
7. Shostack, A.: Threat Modeling: Designing for Security. Wiley (2014)
8. Tete, S.B.: Threat Modelling and Risk Analysis for Large Language Model (LLM)-Powered Applications (2024), `https://arxiv.org/abs/2406.11007`

# Towards Session-Aware Kubernetes: Initial Approach for AR Telepresence

Simão Santos[1] and Nuno Pereira[1]

INESC TEC and Instituto Politécnico do Porto, IPP, Portugal
simao.p.santos@inesctec.pt, nuno.a.pereira@inesctec.pt

**Abstract.** Integrating Augmented Reality (AR) with telepresence enables immersive remote collaboration by virtually placing distant users in shared environments. However, AR workloads often exceed the capabilities of most client devices, necessitating computational offloading to Edge or Cloud infrastructure. These offloaded processes are highly latency-sensitive and require session-aware orchestration across distributed nodes. Kubernetes, the de facto standard platform for managing workloads in cloud environments, is designed primarily for stateless, REST-based applications and lacks the inherent capability to handle AR telepresence systems' dynamic, session-centric nature.

This work proposes an initial design to extend Kubernetes with native support for telepresence sessions. Central to our design is a custom resource called "Session", which enables the platform to understand and manage complete session life cycles. This approach is implemented through a dedicated Session Controller that encapsulates the orchestration logic, dynamically reconciling workloads as users join or leave sessions. By introducing session awareness at the platform level, our solution provides the specialized management capabilities required for latency-sensitive AR telepresence applications while building upon the widely adopted Kubernetes ecosystem.

**Keywords:** Telepresence · Kubernetes · Edge Computing.

## 1 Introduction

While Kubernetes [6] has become the de facto standard for orchestrating containerized applications, its native support for AR telepresence workloads remains limited. Unlike traditional REST-based applications, telepresence involves session-based interactions – dynamic groups of users engaging within a shared virtual environment. Applications that incorporate AR functionality are typically resource-intensive, often surpassing the processing power and battery capacity of end-user devices [11, 17, 19, 26]. To handle tasks such as object detection and high-resolution rendering, these applications require offloading computation to remote servers [15, 26].

Moreover, AR systems are highly sensitive to latency, demanding end-to-end delays of less than 15 milliseconds to ensure a seamless user experience [12, 24]. Achieving such low latency necessitates executing computations close to end-users, typically through distributed edge infrastructure [15, 22]. While some studies have investigated using Kubernetes for Extended Reality (XR) applications, they have not addressed the session-centric requirements fundamental to telepresence [12, 25].

This work's objective is to explore an initial design to extend Kubernetes to enable native support for session-based, latency-sensitive workloads such as AR telepresence. Specifically, we identify the following key requirements:

– Extend Kubernetes with a native workload management resource capable of orchestrating session-based applications.
– Enable the deployment of workloads across geographically distributed infrastructure (i.e., multiple clusters) to minimize latency.
– Provide a unified interface that abstracts the complexity of multi-cluster environments, enabling application runtime to deploy and manage workloads as users join or leave sessions.

The remainder of this paper is organized as follows. Section 2 reviews relevant related work. Section 3 presents an initial design that extends Kubernetes to meet the above requirements, enabling native support for session-based latency-sensitive workloads. Finally, in Section 4, we present some conclusions and future work.

## 2 Related Work

Kubernetes is an open source system designed to automate the deployment, scaling, and management of containerized applications. At its core, Kubernetes operates on a declarative model, allowing users to specify a desired system state. This desired state is continuously maintained by control loops that work to align the actual cluster state with the declared configuration. For an in-depth understanding of Kubernetes' inner workings, the reader is encouraged to consult the official documentation.

Kubernetes was originally designed for cloud-centric scenarios, and prior studies have highlighted a certain degree of unsuitability in edge environments due to a specific design decision: it stores the entire cluster state in etcd, a strongly consistent key-value store [18]. Etcd, a control-plane component that serves as the Kubernetes cluster's single source of truth, stores information about the desired cluster state, the current status of applications, nodes, and other resources [18]. Horizontally scaling this component sacrifices availability in favor of consistency, illustrating the CAP theorem's trade-off between these two properties. State-of-the-art solutions propose multi-cluster Kubernetes deployments instead of single-cluster setups to achieve better scalability [12, 20, 23, 25]. While this approach overcomes the limitations of cluster size, it introduces two new challenges: multi-cluster management and inter-cluster networking.

### 2.1 Multi-Cluster Management

Cluster API (CAPI) [3] is an open-source, vendor-agnostic Kubernetes sub-project that simplifies the coordination and deployment of multiple Kubernetes clusters. It enables cluster definitions through YAML files, facilitating the automation of each cluster life-cycle (create, upgrade and delete). Simply put, CAPI provides a set of components that enable the creation of a Kubernetes management cluster designed specifically to manage the life-cycle of other clusters handling actual application workload. Previous work has proposed the utilization of CAPI in the context of multi-cluster management. Theodoropoulos et al. [25] introduced a multi-cluster orchestration framework designed to support XR services, leveraging CAPI as a key

architectural enabler. The framework comprises a central management cluster, which hosts CAPI components and intelligent algorithms, along with multiple workload Kubernetes clusters responsible for executing XR services. Similarly, Nguyen et al. [20] utilized CAPI to abstract a set of Kubernetes clusters deployed across heterogeneous infrastructures and vendors into a single management cluster, offering an approach to managing multiple Kubernetes clusters as if they were a single entity.

### 2.2 Inter-Cluster Networking

Deploying Kubernetes clusters across various edge and cloud sites often results in a fragmented network infrastructure, which can limit communication between services deployed in separate clusters. Previous work has addressed the challenge of establishing cross-cluster networking by adopting networking technologies that create virtual networks across multiple clusters [12–14, 16, 20, 21, 23, 25]. These include Cilium Cluster [2], Submariner [10] and Liqo [10]. Collectively, these approaches solve the challenge of enabling communication between application services deployed in different clusters.

### 2.3 Custom Workload Management Resources

Beyond the challenges posed by multi-cluster environments, such as management and networking discussed above, a specific issue arises in telepresence scenarios: the need to manage workloads dynamically based on session events. Kubernetes does not natively understand the concept of a "telepresence session" or how to automatically adjust resources as clients join or leave these sessions. Addressing this requires custom workload management capabilities. An illustrative example of extending Kubernetes for specific workload types is Agones [1]. Agones is an open-source platform built on top of Kubernetes designed for orchestrating dedicated game servers for large-scale multiplayer games. Similar to the objective of this work — to provide native Kubernetes management for a specific application domain — Agones enables developers to focus on their game logic rather than the underlying infrastructure. It achieves this by introducing custom resources and controllers that allow Kubernetes to create, run, manage, and scale dedicated game server processes using standard Kubernetes tooling and APIs. This demonstrates how Kubernetes can be extended to handle the lifecycle and scaling of application-specific workloads that do not fit the standard application models.

*Custom Resource Definitions and Operator Pattern* To manage session-based workloads like the ones in telepresence scenarios, Kubernetes can be extended using Custom Resource Definitions (CRDs) [7] and the Operator pattern [8]. A CRD allows the definition of custom API objects with user-defined schemas, which the Kubernetes API server [5] exposes as RESTful endpoints. However, CRDs alone are passive data structures; they require custom controllers — known as Operators — to implement the logic that manages these resources.

Operators encapsulate domain-specific knowledge, enabling Kubernetes to control the lifecycle of non-standard workloads through native abstractions [8]. This approach

mirrors how Agones extends Kubernetes for game server orchestration and is similarly applicable to session-based telepresence environments.

While the reviewed literature addresses common challenges to the ones of this work, it does not target the core challenge of orchestrating workloads with session-awareness. The reviewed works, although related to AR application orchestration across cloud-edge environments . For this reason, we intend to build upon current practices and further complement

## 3 Extending Kubernetes

At the core of our proposal to enable native Kubernetes support for session-based Workloads is a new custom resource called Session. The motivation for introducing this resource stems from the need for the cluster to natively understand and manage the concept of a telepresence session. This design philosophy mirrors the way Kubernetes handles its built-in resources. For instance, when creating a Deployment, the Kubernetes API inherently understands the Deployment resource and manages the corresponding workload accordingly. In a similar vein, the goal of this new architectural concept is to extend Kubernetes with native support for telepresence sessions, allowing it to orchestrate related workloads in an integrated manner.

Therefore, this extension requires a controller capable of reconciling objects of this new kind: the Session Controller. Moreover, to handle multiple clusters and abstract the associated complexity, we propose a Session Manager acting as a bridge between the various infrastructure nodes.

### 3.1 Session Controller

The Session Controller is the core component of the proposed solution, encapsulating all the domain-specific logic required to orchestrate telepresence applications. In this architecture, the Kubernetes API is extended through a CRD, which introduces a new resource kind called Session. As a result, Session objects are treated like any other native Kubernetes resource and are persisted in etcd. Each Session object adheres to a defined schema that specifies the pods to be scheduled when clients join the system. Additionally, it maintains a record of the clients participating in the session. This design allows the Session Controller to react to changes in the Session object. For instance, when a new client is registered, the controller reconciles the updated object by launching the appropriate workloads. Conversely, it can scale down or remove workloads as clients leave the session.

### 3.2 Session Manager

The Session Manager acts as a bridge across multiple Kubernetes clusters, consolidating their APIs into a single, unified interface. This abstraction simplifies session management, handling tasks like user registration and deletion. When a new Session is initiated, the Session Manager replicates this resource to all participating clusters. Consequently, users joining a session must designate their preferred cluster

for workload deployment. The Session Manager then registers the client within the chosen cluster and updates the relevant Session object. These modifications are detected by the Session Controller, which reconciles the updated object and initiates the necessary workloads.

### 3.3 Networking

From a networking perspective, each deployed pod requires independent access, as it is designated to accommodate specific users within a session. In Kubernetes, services typically abstract a deployment's pods by load-balancing requests among them. However, this standard model doesn't inherently allow for the direct, individual addressing of each pod. To overcome this limitation, the Session Controller creates a dedicated Kubernetes service for each pod. This approach, similar to examples provided by the Metacontroller [4], ensures that every pod is uniquely accessible. Furthermore, these individual pod services are registered with an ingress resource managed by the Nginx Ingress Controller [9]. Additionally, each cluster features a ping server, allowing users to measure network latency before connecting.

## 4 Conclusion and Future Work

This work aims to address the gap in Kubernetes' native capabilities for managing session-based, latency-sensitive workloads such as AR telepresence applications. These applications require dynamic resource management, low-latency execution, and coordination across geographically distributed infrastructure — challenges that extend beyond Kubernetes' default workload model.

To overcome these limitations, we propose a novel architectural solution that extends Kubernetes with a custom resource, Session, and its associated controller. This design enables Kubernetes to natively interpret and orchestrate telepresence sessions, scaling workloads up or down as users join or leave, and distributing resources across multiple clusters to optimize latency.

The proposed system has been fully implemented and is scheduled for evaluation as part of future work. We intend to measure key performance metrics, including:

- *Resource efficiency:* Comparing resource utilization against Kubernetes' native workload abstractions to assess overhead and waste.
- *Onboarding latency:* Measuring the time required for application services to become available after a user joins a session.
- *Recovery time:* Observing how quickly application services recover from pod failures or restarts.
- *Latency trade-offs:* Analyzing end-to-end delays between cloud and edge sites.
- *Portability:* Experimenting with the system in alternative Kubernetes distributions, such as KubeEdge, to evaluate its performance in edge-native platforms.

## References

1. Agones documentation. `https://agones.dev/site/docs/`, [Accessed 15-05-2025]

2. Cilium documentation. https://docs.cilium.io/en/stable/, [Accessed 15-05-2025]
3. Cluster api documentation. https://clusterapi.sigs.k8s.io/, [Accessed 15-05-2025]
4. Google cloud platform: Metacontroller. https://github.com/GoogleCloudPlatform/metacontroller/blob/master/examples/service-per-pod/README.md, [Accessed 15-05-2025]
5. Kubernetes api reference. https://kubernetes.io/docs/reference/generated/kubernetes-api/v1.32/, [Accessed 15-05-2025]
6. Kubernetes documentation. https://kubernetes.io/, [Accessed 15-05-2025]
7. Kubernetes documentation: Extend the kubernetes api with customresourcedefinitions. https://kubernetes.io/docs/tasks/extend-kubernetes/custom-resources/custom-resource-definitions/, [Accessed 15-05-2025]
8. Kubernetes documentation: Operator pattern. https://kubernetes.io/docs/concepts/extend-kubernetes/operator/, [Accessed 15-05-2025]
9. Nginx controller documentation. https://docs.nginx.com/nginx-ingress-controller/, [Accessed 15-05-2025]
10. Submariner documentation. https://submariner.io/, [Accessed 15-05-2025]
11. Bartolomeo, G., Cao, J., Su, X., Mohan, N.: Characterizing distributed mobile augmented reality applications at the edge. In: CoNEXT Companion 2023 - Companion of the 19th International Conference on emerging Networking EXperiments and Technologies. pp. 9–18. Association for Computing Machinery, Inc (12 2023). https://doi.org/10.1145/3624354.3630584
12. Benmerar, T.Z., Theodoropoulos, T., Fevereiro, D., Rosa, L., Rodrigues, J., Taleb, T., Barone, P., Giuliani, G., Tserpes, K., Cordeiro, L.: Towards establishing intelligent multi-domain edge orchestration for highly distributed immersive services: a virtual touring use case. Cluster Computing **27**, 4223–4253 (7 2024). https://doi.org/10.1007/s10586-024-04413-7
13. Bringhenti, D., Sisto, R., Valenza, F.: Security automation for multi-cluster orchestration in kubernetes. In: 2023 IEEE 9th International Conference on Network Softwarization: Boosting Future Networks through Advanced Softwarization, NetSoft 2023 - Proceedings. pp. 480–485. Institute of Electrical and Electronics Engineers Inc. (2023). https://doi.org/10.1109/NetSoft57336.2023.10175419
14. Chiaro, C., Monaco, D., Sacco, A., Casetti, C., Marchetto, G.: Latency-aware scheduling in the cloud-edge continuum. In: Proceedings of IEEE/IFIP Network Operations and Management Symposium 2024, NOMS 2024. Institute of Electrical and Electronics Engineers Inc. (2024). https://doi.org/10.1109/NOMS59830.2024.10575183
15. Cozzolino, V., Tonetto, L., Mohan, N., Ding, A.Y., Ott, J.: Nimbus: Towards latency-energy efficient task offloading for ar services. IEEE Transactions on Cloud Computing **11**(2), 1530–1545 (2023). https://doi.org/10.1109/TCC.2022.3146615
16. Ejaz, S., Al-Naday, M.: Fork: A kubernetes-compatible federated orchestrator of fog-native applications over multi-domain edge-to-cloud ecosystems. In: Proceedings of the 27th Conference on Innovation in Clouds, Internet and Networks, ICIN 2024. pp. 57–64. Institute of Electrical and Electronics Engineers Inc. (2024). https://doi.org/10.1109/ICIN60470.2024.10494435
17. Guo, Y., Zou, B., Ren, J., Liu, Q., Zhang, D., Zhang, Y.: Distributed and efficient object detection via interactions among devices, edge, and cloud. IEEE Transactions on Multimedia **21** (2019). https://doi.org/10.1109/TMM.2019.2912703
18. Jeffery, A., Howard, H., Mortier, R.: Rearchitecting kubernetes for the edge. In: EdgeSys 2021 - Proceedings of the 4th International Workshop on Edge Systems, Analytics and Networking, Part of EuroSys 2021 (2021). https://doi.org/10.1145/3434770.3459730
19. Liu, L., Li, H., Gruteser, M.: Edge assisted real-time object detection for mobile augmented reality. In: Proceedings of the Annual International Conference on Mobile Computing and Networking, MOBICOM (2019). https://doi.org/10.1145/3300061.3300116
20. Nguyen, T.N., Lee, J., Vitumbiko, M., Kim, Y.: A design and development of operator for logical kubernetes cluster over distributed clouds. In: Proceedings of IEEE/IFIP Network Operations and Management Symposium 2024, NOMS 2024. Institute of Electrical and Electronics Engineers Inc. (2024). https://doi.org/10.1109/NOMS59830.2024.10575914
21. Poggiani, L., Puliafito, C., Virdis, A., Mingozzi, E.: Live migration of multi-container kubernetes pods in multi-cluster serverless edge systems. In: Proceedings of the 1st Workshop on Serverless at the Edge. pp. 9–16. Association for Computing Machinery (ACM) (2024). https://doi.org/10.1145/3660319.3660330
22. Siriwardhana, Y., Porambage, P., Liyanage, M., Ylianttila, M.: A survey on mobile augmented reality with 5g mobile edge computing: Architectures, applications, and technical aspects (2021). https://doi.org/10.1109/COMST.2021.3061981

23. Syrigos, I., Makris, N., Korakis, T.: Multi-cluster orchestration of 5g experimental deployments in kubernetes over high-speed fabric. In: 2023 IEEE Globecom Workshops, GC Wkshps 2023. pp. 1764–1769. Institute of Electrical and Electronics Engineers Inc. (2023). https://doi.org/10.1109/GCWkshps58843.2023.10465054
24. Theodoropoulos, T., Makris, A., Boudi, A., Taleb, T., Herzog, U., Rosa, L., Cordeiro, L., Tserpes, K., Spatafora, E., Romussi, A., Zschau, E., Kamarianakis, M., Protopsaltis, A., Papagiannakis, G., Dazzi, P.: Cloud-based xr services: A survey on relevant challenges and enabling technologies. Journal of Networking and Network Applications **2** (2022). https://doi.org/10.33969/j-nana.2022.020101
25. Theodoropoulos, T., Rosa, L., Boudi, A., Benmerar, T.Z., Makris, A., Taleb, T., Cordeiro, L., Tserpes, K., Song, J.S.: Cross-cluster networking to support extended reality services. IEEE Network (2024). https://doi.org/10.1109/MNET.2024.3453301
26. Zhang, L., Wu, X., Wang, F., Sun, A., Cui, L., Liu, J.: Edge-based video stream generation for multi-party mobile augmented reality. IEEE Transactions on Mobile Computing **23**(1), 409–422 (2024). https://doi.org/10.1109/TMC.2022.3232543

# A Review on the Use of Large Language Models in Threat Model Generation

Ana Batista[1], Pedro Pinto[1,2], and Nuno Pereira[1,2]

[1] Instituto Superior de Engenharia do Porto, Instituto Politécnico do Porto, IPP, Portugal
[2] INESC TEC, Porto, Portugal
{1231416, pfp, nap}@isep.ipp.pt

**Abstract.** Threat modeling is a crucial task during software development; however, it requires time, specialized knowledge, and domain-specific expertise. Large Language Models (LLMs) have gained significant popularity and are emerging as promising tools to support or even automate this cybersecurity task. This paper presents a literature review to assess the current state of research in this domain using the PRISMA methodology, through which three recent studies were identified. As new LLM-based threat modeling tools arise, the need for effective evaluation of their outputs becomes crucial, motivating the development of a systematic evaluation framework for tasks involving automated threat identification. LLM-as-a-judge is positioned as an innovative technique to be integrated into this framework, testing the capabilities of LLMs to review and score threats generated by other models.

**Keywords:** Threat Modeling · Large Language Models · LLM-as-a-judge

## 1 Introduction

Threat modeling is recognized as a fundamental practice in the early stages of software development, particularly during the planning phase. The insights derived from this process serve as critical security guidelines that inform decision-making throughout the software development lifecycle. Within the Development, Security, and Operations (DevSecOps) framework, threat modeling plays an essential role by enabling a more structured approach to continuous security assessment [5].

Despite its importance, implementing threat modeling can be time-consuming and complex, especially for development teams that lack specialized cybersecurity expertise. This challenge has led to growing interest in alternative approaches aimed at streamlining or automating parts of the process. Among these, Artificial Intelligence (AI), particularly Large Language Models (LLMs), has emerged as a promising solution. As LLMs advance in capability, they are increasingly adopted across various industries to tackle domain-specific problems. However, their application to threat modeling remains underexplored, and existing literature offers limited insight into their practical effectiveness in this context [6]. A systematic review of the current state of the art is therefore necessary to assess the landscape and determine how LLMs can contribute to or enhance traditional threat modeling practices.

This paper adopts the PRISMA methodology to identify and analyze peer-reviewed studies and research initiatives focused on the use of LLMs for threat identification. The purpose of this review was to uncover key trends and highlight areas that warrant further investigation.

Given the varied approaches to model evaluation and the limitations observed across the three selected studies, an interesting area for research unfolds. Since LLM outputs are highly sensitive to prompt design and other performance techniques, assessing and comparing their quality requires multiple approaches, including well-defined evaluation criteria. Therefore, identifying suitable evaluation metrics is essential. The adoption of systematic evaluation frameworks is proposed as a necessary step toward ensuring that LLM-based threat modeling tools can be reliably used for the task.

The document is structured as follows: Section 2 addresses ethical considerations. Section 3 describes the steps taken to organize the review in accordance with the agreed-upon methodology. Section 4 presents the findings derived from the reviewed articles. Section 5 discusses these findings and explores potential directions for future work. Finally, Section 6 concludes the paper with final remarks.

## 2 Ethical Considerations

Generative AI models, such as ChatGPT and Gemini, can significantly lower the barriers of technical expertise and time commitment by autonomously generating relevant threat scenarios.

However, these models carry inherent limitations (e.g., hallucinations) [3], as well as legal and ethical implications. On March 20th, 2023, a data breach involving ChatGPT exposed user conversations, violating the privacy of many individuals. With similar consequences, these models state that they may use user conversations with the goal of improving their solutions, which raises additional privacy concerns [2].

Therefore, despite their advantages in generating threat models, their impact on information confidentiality and integrity can impact the level of trust placed in them for such tasks.

## 3 Methodology

This review explores how the creation of threat models through AI has been approached, identifying its applications and respective advantages and disadvantages, while essentially analyzing how these developments are transforming contemporary practices in cybersecurity.

Thus, the following two research questions were formulated:

- RQ1: What drives the use of LLMs to generate threat models instead of relying on manual procedures?
- RQ2: Do the techniques integrated into LLMs enhance the performance of the original models to produce higher-quality threat modeling?

Then, the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) methodology [8] was used to organize the research and synthesize the results obtained. To adhere to PRISMA guidelines, keywords must first be categorized to target and focus on answering the defined research questions.

Table 1 presents the query used, employing Boolean search operators such as "AND" and "OR" to combine the keywords LLM and threat modeling, and refine the search

within the selected databases. The table includes the inclusion and exclusion criteria. Only studies from the past five years are considered, given the recent emergence of LLM-based chatbots. Furthermore, studies will be excluded if they lack structure and proper referencing, are not written in English, or do not focus on supporting threat identification through AI.

**Table 1.** Query and Selection Criterion

| **Search Query** | **Searching in Databases (until May, 2025)** | **Inclusion Criteria** | **Exclusion Criteria** |
|---|---|---|---|
| LLM AND "threat modeling" | **ScienceDirect** uses the "Title, abstract, or author-specified keywords" field. **IEEE** lacks this field and uses Command Search instead: *(("Abstract":docker OR "Document Title":LLM OR "Author Keywords":LLM) AND ("Abstract":"threat modeling" OR "Document Title":"threat modeling" OR "Author Keywords":"threat modeling"))* **ArXiv** does not offer any of these filtering options, so the search was conducted across all fields. | 1) Studies from 2020–2025, since it is still an emerging area 2) Studies in English 3) Studies on the use of Artificial Intelligence to support and/or automate threat modeling | 1) Studies not in English 2) Articles that are not peer-reviewed or lack proper formatting 3) Studies without a focus on threats identification and their generation using LLMs |

Fig. 1 presents the stages and number of papers processed in this systematic review. The process began with the identification stage, where 69 papers were retrieved from academic databases using the search query defined, along with filters such as publication year (2020 – 2025) and keyword presence in the title, abstract, or author-defined fields. A full-text screening of these 69 articles led to the exclusion of studies based on exclusion criterion 3: studies addressing cybersecurity in general without a specific focus on threat models or their generation through LLMs As a result, 66 articles were excluded. During the inclusion stage, a final selection of 3 studies was made, all of which demonstrated high relevance to the research objectives.

## 4 Results and Analysis

Grounded in the findings obtained through the PRISMA-identified studies, the following two subsections aim to address the research questions previously introduced. Subsection 3.1 explores RQ1: What drives the use of LLMs to generate threat models instead of relying on manual procedures? The subsection 3.2 addresses RQ2: Do the techniques integrated into LLMs enhance the performance of the original models to produce higher-quality threat modeling?

### 4.1 The Adoption of LLMs for Generating Threat Models

Among the projects relevant to answering RQ1, one notable example is STRIDE GPT, developed by Matt Adams, designed to harness the capabilities of an LLM for

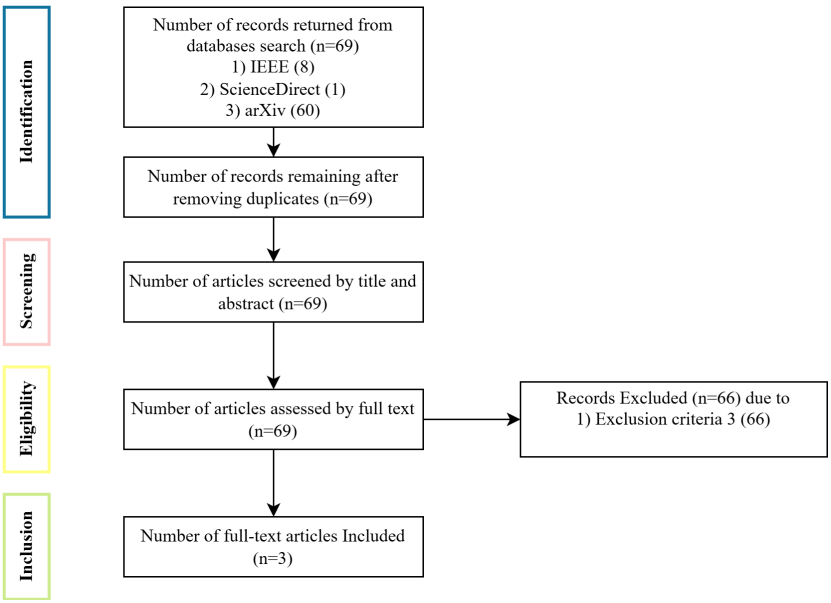



**Fig. 1.** PRISMA Diagram

cybersecurity applications [1]. Referenced by [9], STRIDE GPT is used to generate an initial threat model with Microsoft's STRIDE - Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, and Elevation of Privilege. Incorporating this tool involves analyzing a web application, breaking it down into its components and services, and leveraging AI to automatically assess the potential impact of each category on the overall application, based on the available information. STRIDE GPT can use existing AI models, including the latest GPT models [1].

Cyber Sentinel emerges as a conversational agent designed to optimize cybersecurity. It is not only aimed at assisting in Threat Modeling but also supports various other cybersecurity operations. Transforms conversations into an interactive and collaborative method for discussing and sharing knowledge with security analysts, while also raising awareness among employees [4].

One issue with STRIDE GPT, as analyzed by [9], is that it may make errors in organizing threats into the correct categories. If prompt engineering is not well-studied and processed, the models may interpret the request differently each time, leading to inconsistent classifications. Cyber Sentinel, although highly adaptable to new threats, struggles to provide effective mitigation strategies.

ThreatModeling-LLM is a framework created with the same objective of automating threat identification and mitigation suggestions, specifically for banking systems, overcoming the limitations of STRIDE GPT and Cyber Sentinel [9].

Similarly, the study conducted by [6] aimed to fine-tune and optimize the Llama2 model for Threat Modeling in the domain of modern medical devices – the CyberLlama2 model.

### 4.2 Optimizing Approaches to Enhance LLMs for Threat Modeling

As demonstrated by the author of ThreatModeling-LLM in [9], the integration of fine-tuning and prompt engineering strategies significantly improves the performance of LLMs. Fine-tuning is essential, as the models are not initially equipped to answer cybersecurity-related questions [6], while prompt engineering addresses how the questions are formulated. Whether complex or ambiguous, the formulation of questions can influence the model's ability to interpret them effectively [4]. Table 2 summarizes the techniques applied for Cyber Sentinel, ThreatModeling-LLM, and CyberLlama2 discussed in the previous subsection. This analysis offers a clear and evidence-based response to RQ2, highlighting the necessity of iterative experimentation and optimization to determine the most effective configuration.

The authors employed some prompt engineering techniques, particularly in Cyber Sentinel and ThreatModeling-LLM, to enhance model performance. Among these, Chain of Thought (CoT) prompting emerged as the commonly used strategy, enabling more accurate threat assessments through step-by-step reasoning. In addition, ThreatModeling-LLM combined prompt engineering with fine-tuning to better tailor the model to the specific demands of banking systems, while CyberLlama2 relied primarily on fine-tuning to adapt the model to the medical domain. ThreatModeling-LLM demonstrated that the combination of prompt engineering and fine-tuning yielded the most effective results. STRIDE GPT is expected to represent a subject of future study, as the specific strategies it employs remain unclear and require further investigation.

**Table 2.** Techniques for Enhancing Models' Performance

| | **Cyber Sentinel (GPT-4 models)** | **ThreatModeling-LLM (Llama-3.1-8B and GPT-3.5-turbo models)** | **CyberLlama2 (Llama2 model optimization)** |
|---|---|---|---|
| **Objective** | Serve as an intelligent conversational assistant to support and validate threat models, among other cybersecurity tasks | Automate the threat modeling process in banking systems based on the STRIDE framework and NIST 800-53 control codes | Automate the threat modeling process in the Medical Field (MMD) based on the MEDICALHARM methodology |
| **No techniques applied** | An initial prompt is given to the model to identify the type of request made by the user | Values of metrics: Accuracy (0.17), Precision (0.35), and Recall (0.27) | Llama2 baseline model was evaluated using metrics such as ROUGE, MAUVE, and METEOR |
| **Fine-tuning** | Not applicable | Through fine-tuning, the models achieved over 60% accuracy, and the other metrics (precision, recall, and text similarity with BERT) also improved | Fine-tuning the Llama2 model with data from various cybersecurity sources. 2,000 entries were tested, achieving better results in most metrics compared to the original Llama2 model |

| | Cyber Sentinel (GPT-4 models) | ThreatModeling-LLM (Llama-3.1-8B and GPT-3.5-turbo models) | CyberLlama2 (Llama2 model optimization) |
|---|---|---|---|
| **Zero-shot prompting** | Not applicable | Not applicable | Enabling the model to respond to new questions based on previously acquired knowledge |
| **Chain-of-Thought (CoT)** | Through this technique, the model provides additional context in steps | Accuracy and Recall increased, showing that articulated reasoning processes help improve the identification of threats | Not applicable |
| **Optimization by Prompting (OPRO)** | Not applicable | Precision achieves moderate improvements, but does not reach CoT's overall performance | Not applicable |
| **Chain-of-Thought + OPRO** | Not applicable | Achieved the highest scores in all metrics. Precision and Recall almost reached 0.6, and Text Similarity, measured with BERT, exceeded 0.95 | Not applicable |
| **Self-Consistency** | GPT-4 generates approaches and compares them to choose the most appropriate one | Not applicable | Not applicable |

## 5 Discussion

The evaluation of the results proved essential in justifying the performance techniques employed (RQ2) and, consequently, in assessing the feasibility of using LLM-based threat modeling tools for the task of generating threat models (RQ1). As shown in Table 2, the metrics adopted across studies range from traditional approaches, such as ROUGE and METEOR, to LLM-enhanced methods, including semantic similarity metrics like BERTScore, complemented by expert evaluations.

A closer examination of these metrics reveals that many may fall short in effectively capturing the specific requirements of the threat model generation task. Traditional metrics primarily assess lexical similarity, failing to account for contextual meaning, an essential factor in tasks of this nature [7]. BERTScore approach represents a significant improvement, as noted by [7], yet still presents certain limitations. Therefore, careful selection and comparison of evaluation metrics becomes a matter of critical importance.

In light of these considerations, this discussion advocates for the development of a systematic evaluation framework specifically designed for this task, incorporating diverse evaluation strategies. While traditional metrics and semantic similarity measures remain essential for inclusion and comparison, their known shortcomings highlight the need for complementary approaches.

G-Eval, introduced by [7], acts as an LLM-as-a-Judge that harnesses the capabilities of LLMs to evaluate outputs generated by other AI systems. When provided with a clear explanation of the task and its evaluation criteria (dimensions)

through a well-crafted user-defined prompt, the LLM generates a detailed evaluation process, an automatic Chain-of-Thought, and completes a structured scoring form for each criterion, ultimately producing a final score based on probabilistic reasoning.

As an approach towards an evaluation framework, the evaluation dimensions relevant to threat modeling tasks may include:

- Consistency: How well does the generated threat model align with the reference in terms of factual accuracy? Are the threats similar, and if there are any additional threats, are they justified based on the threats in the reference?
- Relevance: Assesses whether the identified threats are significant and pertinent to the context, avoiding unnecessary repetition or inclusion of low-value information. Each threat should provide meaningful insight rather than serve as filler.
- Coverage: Examines the extent of the generated model in relation to the reference, checking if important and contextually relevant threats have been covered.
- Plausibility: Do the threats make technical sense given the system's inferred characteristics? This evaluation criterion assesses whether the threats are logically coherent and free from technical inaccuracies or contradictions.

Thus, the different approaches, ranging from traditional methods to more innovative LLM-based solutions, are primarily designed to reinforce the trust developers have in LLMs to assist in securing their systems. Integrating these approaches into a systematic evaluation framework aims to collect meaningful insights and further support this objective.

## 6 Conclusion

The growing interest in LLMs has sparked extensive discussions regarding their potential applications in cybersecurity.

This paper analyzes three studies selected using the PRISMA methodology, focusing on the effectiveness of LLMs in threat identification. The findings indicate that the performance of these models is largely influenced by prompt engineering and fine-tuning. Despite these advancements, a critical challenge has emerged from the analysis: how to determine the most appropriate metrics and accurately evaluate the outputs of model-generated threat models.

A proposed future implementation involves the development of a systematic evaluation framework specifically designed to assess threat models generated by LLMs. This framework aims to identify the most effective metrics, compare them, and strike a balance between automated results and human validation, ensuring that the outcomes are helpful. Ultimately, it will assist developers in this essential process, empowering them to consistently analyse the security of their systems.

## References

1. Adams, M., Shibata, K.: Stride gpt: An ai-powered threat modeling tool. `https://github.com/mrwadams/stride-gpt` (2024)

2. Chowdhury, M.M., Rifat, N., Ahsan, M., Latif, S., Gomes, R., Rahman, M.S.: Chatgpt: A threat against the cia triad of cyber security. In: 2023 IEEE International Conference on Electro Information Technology (eIT). pp. 1–6 (2023). https://doi.org/10.1109/eIT57321.2023.10187355
3. Gupta, M., Akiri, C., Aryal, K., Parker, E., Praharaj, L.: From chatgpt to threatgpt: Impact of generative ai in cybersecurity and privacy. IEEE Access **11**, 80218–80245 (2023). https://doi.org/10.1109/ACCESS.2023.3300381
4. Kaheh, M., Kholgh, D.K., Kostakos, P.: Cyber sentinel: Exploring conversational agents in streamlining security tasks with gpt-4 (2023), https://arxiv.org/abs/2309.16422
5. Kushwaha, M.K., David, P., Suseela, G.: Automation and devsecops: Streamlining security measures in financial system. In: 2024 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT). pp. 1–6 (2024). https://doi.org/10.1109/CONECCT62155.2024.10677271
6. Kwarteng, E., Cebe, M., Kwarteng, J.: Cyberllama2 - medicalharm threat modeling assistant. In: 2024 International Conference on Machine Learning and Applications (ICMLA). pp. 930–934 (2024). https://doi.org/10.1109/ICMLA61862.2024.00136
7. Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., Zhu, C.: G-eval: Nlg evaluation using gpt-4 with better human alignment (2023), https://arxiv.org/abs/2303.16634
8. Page, M.J., et al.: The prisma 2020 statement: an updated guideline for reporting systematic reviews. BMJ **372** (2021). https://doi.org/10.1136/bmj.n71, https://www.bmj.com/content/372/bmj.n71
9. Yang, S., Wu, T., Liu, S., Nguyen, D., Jang, S., Abuadbba, A.: Threatmodeling-llm: Automating threat modeling using large language models for banking system (2024), https://arxiv.org/abs/2411.17058

# Initial Explorations in Industrial Video Summarization with LLMs and MLLMs

Rui Neto, Nuno Pereira and Paula Viana

INESC TEC, Porto, Portugal and ISEP, Polytechnic of Porto, Porto, Portugal
{rui.j.neto, nuno.a.pereira, paula.viana}@inesctec.pt {1230211, nap, pmv}@isep.ipp.pt

**Abstract.** In industrial settings, operational safety and efficiency often rely on accurately detecting machine-related anomalies. Recent advancements in artificial intelligence, particularly with large language models (LLMs) and multi-modal large language models (MLLMs), have the potential to revolutionize video understanding and summarization for these critical tasks. However, their application in structured video summarization tailored for industrial contexts remains largely unexplored. Key challenges include adapting LLMs and MLLMs to handle domain-specific tasks, such as comprehending human-machine interactions and machine states. Towards designing a pipeline tailored for industrial video data, we review the state-of-the-art techniques and methodologies for applying LLMs and MLLMs to structured video understanding. Specifically, the work analyzes critical steps such as video decoding, frame selection, object detection, frame captioning, temporal awareness strategies, and integration of external content to enrich summaries. We also examine and explore evaluation metrics and datasets relevant to industrial scenarios.

**Keywords:** Large Language Models · Multi-modal Large Language Models · Video Summarization · Industrial Environment

## 1 Introduction

In many industries, operational disruptions caused by undetected errors or inefficiencies can lead to safety incidents and financial losses. For example, in 2018, in Japan, 127.329 people were injured in industrial accidents [9]. Companies that actively minimize human errors and workplace incidents, consistently achieve significant long-term cost savings by reducing expenses related to accidents [5]. But training personnel to monitor these events consistently and accurately is not only expensive but also fails to eliminate errors due to fatigue and oversight. Automated detection of machine-related issues in industrial contexts can be an effective way to improve operational safety. For manufacturing and insurance companies, using factory-installed video capture devices to analyze operational data could mean the difference between seamless operations and costly downtime.

Recent deep learning models have revolutionized a various number of industries, including video surveillance [15]. Our research seeks to leverage the capabilities of Large Language Models (LLMs) and Multimodal Large Language Models (MLLMs) to create structured video summaries adapted to industrial applications. These summaries are designed to capture the logical sequence of events and highlight anomalies. In Section 2, we present a systematic literature review, identifying relevant case studies and datasets in the field. Based on these studies, a structured video summarization framework is proposed, detailing key pipeline components such as frame sampling, frame-by-frame labeling, object detection, external content integration, and temporal reasoning. The

work then introduces a methodology for evaluating candidate models. Finally, we discuss current results and outlines future directions for refining and validating the proposed system.

## 2 Related Work

Developing a robust video summarization system hinges on understanding existing approaches and their implementation in similar contexts. Therefore, we adopted a systematic research and delved into notable studies to uncover the steps and technologies that have effectively addressed the challenges of structured video summarization.

### 2.1 Research Method

The current section details the systematic research approach undertaken to identify and analyze relevant literature. The findings provide a foundation for the subsequent analysis of case studies and their steps in industrial video summarization.

**Research Questions** To guide the systematic review, the following research questions were formulated:

- RQ-1: What are the methods for applying LLMs and MLLMs to video summarization tasks, such as video captioning and future event prediction?
- RQ-2: Do existing pre-trained models performing approaches such as few-shot video-language processing, unsupervised language-guided summarization, and zero-shot video understanding help circumvent the need to train LLMs and MLLMs?
- RQ-3: What datasets exist for industrial human action recognition and machine operation validation?

**Data Collection Process** The data collection process closely followed the PRISMA guidelines to ensure a systematic and transparent review. The process involved three key steps:

1. Identification: conducting searches using predefined queries in Google Scholar.
2. Screening: filtering the results based on titles and abstracts to exclude irrelevant and duplicated studies; and assessing the full text of selected studies against the inclusion and exclusion criteria.
3. Inclusion: finalizing the selection of studies for analysis and inclusion in the review.

### 2.2 Case Studies

The research method resulted in five studies, as presented in Table 1. They approach video summarization as a structured pipeline, which closely matches the methodological requirements of our project.

The MISAR system [2] uses AR to enhance task performance through multi-modal integration of visual, auditory, and linguistic data. The core innovation lies in its use of

**Table 1.** Studies chosen for pipeline steps extraction. We assigned the name AVIWT for ease of reference throughout this document, as the original authors did not provide an official name.

| Paper Title | Framework Name | Source | Year |
|---|---|---|---|
| MISAR: a Multi-modal Instructional System with Augmented Reality | MISAR | [2] | 2023 |
| Language Models with Image Descriptors are Strong Few-Shot Video-Language Learners | VidIL | [17] | 2022 |
| M3SUM: A Novel Unsupervised Language-guided Video Summarization | M3SUM | [16] | 2024 |
| A Video Is Worth 4096 Tokens: Verbalize Videos to Understand Them in Zero Shot | AVIWT | [1] | 2023 |
| XaiR: An XR Platform that Integrates Large Language Models with the Physical World | XaiR | [14] | 2024 |

an LLM (GPT-3.5-Turbo) to process and integrate data streams, enabling contextual understanding and real-time user assistance. This approach is particularly aimed at task quantification and error correction in AR environments.

VidIL [17] is a system that uses large-scale pre-trained language models alongside image-language models to perform video-to-text tasks with minimal supervision. The approach is particularly suited for few-shot learning scenarios, requiring no pre-training or finetuning on video datasets. VidIL excels in tasks such as video captioning and question answering, by utilizing a representation of videos that integrates spatial and temporal information.

M3SUM [16] introduces an unsupervised approach to language-guided video summarization, which avoids the need for training data and complex feature alignment models. The system utilizes off-the-shelf models to convert multi-modal video data (frames and audio) into textual descriptions, which are then processed by an LLM for generating video summaries based on the highest score frames.

AVIWT [1] introduces a novel framework that verbalizes videos into coherent textual stories using multi-modal input (visual, audio, and metadata) to enable downstream video understanding tasks in a zero-shot setting. All textual elements, such as transcripts, captions, Optical Character Recognition (OCR) data, and metadata, are combined into a structured prompt for an LLM, GPT-3.5.

XaiR [14] is a platform that integrates MLLMs with XR to enable intelligent task guidance. XaiR facilitates step-by-step task instruction, object detection, and AR content placement. The system is built on a split-architecture design, where heavy computational tasks are offloaded to a server while XR headsets handle real-time spatial interactions.

Through the previous analysis, the key features were identified. These are summarized in Table 2.

### 2.3 Datasets

While deep learning has made significant advances in certain specialized areas, such as medical imaging for cancer detection, many industrial and technical applications

**Table 2.** Comparison of features across different approaches.

| **Feature** | MISAR | VidIL | M3SUM | AVIWT | XaiR |
|---|---|---|---|---|---|
| Frame Sampling | ✔ | ✔ | ✔ | ✔ | ✔ |
| Frame-by-frame Labeling | ✔ | ✔ | ✔ | ✔ | ✔ |
| Object Detection | ✗ | ✔ | ✗ | ✔ | ✔ |
| External Content Integration | ✔ | ✗ | ✗ | ✔ | ✔ |
| Temporal Awareness | ✗ | ✔ | ✔ | ✗ | ✔ |

remain underserved [12]. The challenge lies in the limited availability of comprehensive datasets across various technical domains - from life sciences to manufacturing and industrial processes. This data scarcity has created a notable gap in the practical implementation of Artificial Intelligence (AI) solutions for many specialized technical applications [12]. While no datasets perfectly adapted for the intended purpose were identified, there are some options that offer notable features and can be considered for complementary use: InHARD [4], Assembly101 [13], MECCANO [10], HA4M [3] and HA-ViD, [18]. Simulated environments also offer a cost-effective and flexible alternative for creating datasets that reflect specific real-world scenarios [6]. As a potential solution to the data scarcity, future work could involve constructing a custom dataset by collecting domain-relevant images from publicly available online sources or by generating synthetic data tailored to specific industrial tasks.

## 3 A Video Summarization Framework

Informed by the feature analysis and the specific needs of real-world scenarios, a draft pipeline for video summarization is proposed, depicted in Fig. 1. This pipeline combines the most relevant steps observed in the systems studied.

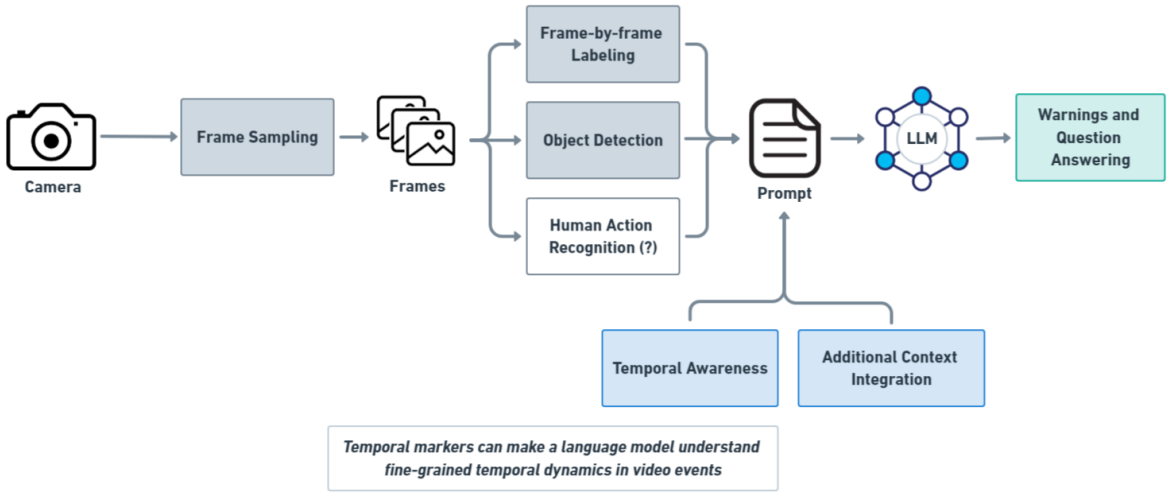



**Fig. 1.** Draft pipeline with the most common steps in video summarization frameworks.

**Frame sampling** captures representative frames of the video while balancing computational efficiency. **Frame-by-frame labeling** generates detailed captions for each sampled frame. **Object detection** enhances frame-level understanding by detecting and labeling significant subjects such as operator, operator and machine features (e.g. body, arms, buttons, levers, gauges). **Human Action Recognition** provides frame-by-frame analysis with the aim of capturing a high-level understanding of the actions performed. **Additional context** and **temporal awareness** enriches the pipeline with metadata, machine details, technical guide information, temporal markers to track dynamics in video, or other relevant machine operation inputs to improve video comprehension. To integrate the information captured across all preceding steps, we use an **LLM** to synthesize them, facilitate anomaly detection, and support advanced tasks such as answering questions. advanced tasks like question answering.

### 3.1 Models and Framework Evaluation

Manual review of each model output and summary by human evaluators provides detailed feedback. Still, it is often impractical due to its high cost, time requirements, and lack of scalability. To address this, automatic evaluation methods are widely employed [11]. These methods aim to approximate the criteria used by human evaluators, focusing on attributes such as fluency, coherence, relevance, factual accuracy, and fairness. Additionally, measuring how closely a generated text aligns in content or style with a reference text is another significant aspect of evaluation [11]. Based on these concepts, we discuss next our proposal to evaluate the models used in the proposed pipeline, in order to identify the most suitable candidate for our framework.

A custom dataset was developed to assess model capabilities in pertinent industrial tasks. We identified five core categories for evaluation: Human-Machine Interaction, Machine State and Identification, Visual Quantification and Estimation, Text and Display Interpretation, and Functional Reasoning. For each category, four distinct questions/tasks were formulated. These tasks were designed to investigate a range of desired model capabilities, including Recognition, Knowledge, OCR, Spatial Awareness, Language Generation, Mathematical Reasoning, and Human Pose Estimation. Using this dataset, we used an LLM-as-a-judge framework to quantitatively assess the results of the model, utilizing the model's capacity to emulate human-like reasoning and cognition [7]. This involves assigning a score (ranging from 0 to 1, with 0.1 intervals) based on the similarity between the ground truth answer from our dataset and the answer generated by the MLLM under evaluation. These scores are then aggregated to derive comprehensive evaluation metrics for each MLLM.

### 3.2 Pipeline Implementation

The major component of our completed work is the construction of an end-to-end processing pipeline. The current pipeline involves several stages.

Frame Sampling and Selection: the input video is subsequently processed by extracting frames every 1 second. To select salient frames, each extracted frame is compared to the previous one using the Structural Similarity Index Measure (SSIM). A frame is selected only if its similarity to the previous frame falls below a predefined threshold.

Object Detection: for each selected frame, object detection is performed using YOLO v11x (executed on an NVIDIA RTX 3090 Ti GPU). The identified objects, along with their corresponding frames and timestamps, are stored in a structured format.

Frame-by-frame Labeling: the selected frames are then sent to a captioning service (currently utilizing Gemini-2.0-Flash, pending final MLLM evaluation). Input to the captioning service includes the frame itself, its timestamp, the objects detected within it, and the captions from the five preceding frames for contextual continuity.

Additional Context Integration with Retrieval-Augmented Generation (RAG): to further enrich the input for summarization, the top-5 relevant entries from the RAG database (using the MongoDB vector store) are retrieved and appended to the information stream.

LLM Summary Generation: the consolidated information (dense captions and RAG context) is fed to the LLM to generate the final video summary.

Preliminary evaluations show that the pipeline functions effectively as a proof of concept, though there remains room for further optimization and refinement.

## 4 Conclusion and Next Steps

This initial exploration aims to provide a foundation to develop systems capable of interpreting complex industrial video data, integrating established computer vision techniques and emerging language model capabilities. By mapping the pipeline components necessary for effective industrial video summarization, we hope to provide a structured foundation for future implementations.

Our planned work encompasses several key areas to refine and validate the proposed framework.

1. Completing the model evaluation to select the best models for integration into the pipeline;
2. Investigating the integration of Human Action Recognition (HAR) techniques to assess their potential value in enhancing summary quality and detail;
3. Studying the LightRAG [8] approach for the additional context integration step.
4. Developing and implementing an LLM-as-a-judge methodology for evaluating the quality of the final generated video summaries;
5. Conducting thorough testing of the entire framework to validate its overall functionality.

## References

1. Bhattacharya, A., Singla, Y.K., Krishnamurthy, B., Shah, R.R., Chen, C.: A video is worth 4096 tokens: Verbalize videos to understand them in zero shot (2023), https://arxiv.org/abs/2305.09758

2. Bi, J., Nguyen, N.M., Vosoughi, A., Xu, C.: Misar: A multimodal instructional system with augmented reality (2023), https://arxiv.org/abs/2310.11699
3. Cicirelli, G., Marani, R., Romeo, L., García Domínguez, M., Heras, J., Perri, A.G., D'Orazio, T.: The ha4m dataset: Multi-modal monitoring of an assembly task for human action recognition in manufacturing. Scientific Data **9**(1), 745 (2022). https://doi.org/10.1038/s41597-022-01843-z, https://doi.org/10.1038/s41597-022-01843-z
4. DALLEL, M., HAVARD, V., BAUDRY, D., SAVATIER, X.: Inhard - industrial human action recognition dataset in the context of industrial collaborative robotics. In: 2020 IEEE International Conference on Human-Machine Systems (ICHMS). pp. 1–6 (2020). https://doi.org/10.1109/ICHMS49158.2020.9209531
5. DeMott, D.L.: Human reliability and the cost of doing business. In: Proceedings of the Annual Maintenance and Reliability Symposium. Society for Maintenance and Reliability Professionals, Galveston, TX, United States (8 2014), https://ntrs.nasa.gov/api/citations/20140008715/downloads/20140008715.pdf, document ID: 20140008715, Report Number: JSC-CN-31348, Acquisition Source: Johnson Space Center
6. Downs, L., Francis, A., Koenig, N., Kinman, B., Hickman, R., Reymann, K., McHugh, T.B., Vanhoucke, V.: Google scanned objects: A high-quality dataset of 3d scanned household items. In: 2022 International Conference on Robotics and Automation (ICRA). pp. 2553–2560 (2022). https://doi.org/10.1109/ICRA46639.2022.9811809
7. Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., Wang, S., Zhang, K., Wang, Y., Gao, W., Ni, L., Guo, J.: A survey on llm-as-a-judge (2025), https://arxiv.org/abs/2411.15594
8. Guo, Z., Xia, L., Yu, Y., Ao, T., Huang, C.: Lightrag: Simple and fast retrieval-augmented generation (2025), https://arxiv.org/abs/2410.05779
9. Hashimoto, S., Ji, Y., Kudo, K., Takahashi, T., Umeda, K.: Anomaly detection based on deep learning using video for prevention of industrial accidents (2020), https://arxiv.org/abs/2005.13734
10. Ragusa, F., Furnari, A., Farinella, G.M.: Meccano: A multimodal egocentric dataset for humans behavior understanding in the industrial-like domain. Computer Vision and Image Understanding **235**, 103764 (2023). https://doi.org/https://doi.org/10.1016/j.cviu.2023.103764, https://www.sciencedirect.com/science/article/pii/S1077314223001443
11. van Schaik, T.A., Pugh, B.: A field guide to automatic evaluation of llm-generated summaries. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 2832–2836. SIGIR '24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3626772.3661346, https://doi.org/10.1145/3626772.3661346
12. Schlagenhauf, T., Landwehr, M.: Industrial machine tool component surface defect dataset. Data in Brief **39**, 107643 (Dec 2021). https://doi.org/10.1016/j.dib.2021.107643, http://dx.doi.org/10.1016/j.dib.2021.107643
13. Sener, F., Chatterjee, D., Shelepov, D., He, K., Singhania, D., Wang, R., Yao, A.: Assembly101: A large-scale multi-view video dataset for understanding procedural activities (2022), https://arxiv.org/abs/2203.14712
14. Srinidhi, S., Lu, E., Rowe, A.: XaiR: An XR Platform that Integrates Large Language Models with the Physical World. In: 2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). pp. 759–767. IEEE Computer Society, Los Alamitos, CA, USA (10 2024). https://doi.org/10.1109/ISMAR62088.2024.00091, https://doi.ieeecomputersociety.org/10.1109/ISMAR62088.2024.00091
15. Suganthi, J., Abineshkumar, V.: Revolutionizing anomaly detection in surveillance footage with arlstm. In: 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC). pp. 506–511 (2023). https://doi.org/10.1109/ICAAIC56838.2023.10141056
16. Wang, H., Zhou, B., Zhang, Z., Du, Y., Ho, D., Wong, K.F.: M3sum: A novel unsupervised language-guided video summarization. In: ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4140–4144 (2024). https://doi.org/10.1109/ICASSP48485.2024.10447504
17. Wang, Z., Li, M., Xu, R., Zhou, L., Lei, J., Lin, X., Wang, S., Yang, Z., Zhu, C., Hoiem, D., Chang, S.F., Bansal, M., Ji, H.: Language models with image descriptors are strong few-shot video-language learners (2022), https://arxiv.org/abs/2205.10747
18. Zheng, H., Lee, R., Lu, Y.: Ha-vid: A human assembly video dataset for comprehensive assembly knowledge understanding. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 67069–67081. Curran Associates, Inc. (2023), https://proceedings.neurips.cc/paper_files/paper/2023/file/d40e6e4b3ee6c24f2bf2cb72c2412f4b-Paper-Datasets_and_Benchmarks.pdf

# Python-Based Tool for Data Cleaning and Validation

Benazir Rostami[1], Inês Sena[2], and Ana I. Pereira[2]

[1] Instituto Politécnico de Bragança, Bragança, Portugal
[2] Research Centre in Digitalization and Intelligent Robotics (CeDRI), Laboratório Associado para a Sustentabilidade e Tecnologia em Regiões de Montanha (SusTEC), Instituto Politécnico de Bragança (IPB), Bragança, Portugal
a61510@alunos.ipb.pt, {ines.sena,apereira}@ipb.pt

**Abstract.** This project presents a Python-based tool designed to automate the data cleaning process for CSV datasets containing user-related records. The tool detects and identifies common data issues such as missing values, invalid entries, and inconsistent formats. It offers flexible correction options: synthetic data generation, manual user correction, or row exclusion. This approach improves the quality and usability of data for downstream tasks like statistical analysis or machine learning.

**Keywords:** Data cleaning · Data validation · Missing values · Synthetic data

## 1 Introduction

In data analysis, the accuracy of results heavily depends on the quality of input data. Raw datasets often have missing values, inconsistent formats, or invalid entries. These issues can significantly skew analytical outcomes or degrade model performance. Hence, data cleaning forms a crucial preliminary step in the data pipeline [3].

Data cleaning improves dataset reliability by identifying and correcting errors, standardizing formats, and handling missing values. High-quality cleaned data supports more accurate models, better visualizations, and sound decision-making [1, 2]. Missing data can arise from incomplete collection, privacy restrictions, or system errors. If unaddressed, it can introduce bias or reduce the representativeness of results. Handling such gaps effectively is key to maintaining data integrity [4, 5].

The main objective of this project was to develop an automated Python-based tool capable of efficiently handling and cleaning CSV files containing user-related data with missing or invalid entries. Rather than manually inspecting thousands of cells, the tool systematically identifies data quality issues, such as missing values, incorrect formats, or invalid fields, and offers users flexible resolution methods, including synthetic data generation, manual correction, or row exclusion. This significantly reduces the time and effort required for data validation and ensures datasets are prepared for accurate analysis or machine learning tasks.

The structure of this paper is as follows: In Section 2, approaches to handling missing data are described. The methodology used to develop the tool is presented in Section 3. The results obtained are presented in Section 4. Finally, Section 6 concludes the study by enumerating possible directions for future research.

## 2 Approaches to Handling Missing Data

Handling missing data is critical in preparing datasets for analysis or Machine Learning (ML) applications [4]. Depending on the nature and extent of the missing values and the downstream goals of the analysis, different strategies can be employed:

- **Imputation Techniques:** Imputation involves replacing missing values with estimated ones, typically the mean, median, or mode. While simple and computationally efficient, this method can artificially reduce data variability and potentially introduce bias [5].
- **Prediction Models:** More advanced methods use machine learning algorithms, such as regression models, decision trees, or k-nearest neighbors, to predict missing values based on the relationships between variables. These approaches generally offer higher accuracy but require more computational resources and assume that the observed data is sufficiently representative [6].
- **Synthetic Data Generation:** This technique replaces missing or invalid entries with synthetic values generated from predefined lists or probabilistic models. This is particularly useful when dealing with personal or sensitive information. Synthetic data enables flexibility and preserves dataset structure, although it must be carefully designed to avoid introducing unrealistic distributions [1].
- **Manual Correction:** Manual input from users or domain experts can correct missing values in smaller datasets or cases where data quality is paramount (e.g., medical or legal records). This method offers high precision but is not scalable for large datasets [3].
- **Deletion:** When a variable or observation has an excessive amount of missing data, it may be more practical to remove it from the dataset. While this simplifies analysis, it risks the loss of valuable information and can decrease the representativeness of the remaining data [5].

## 3 Methodology

This tool was developed using a modular and scalable data validation methodology. The following main steps describe the process:

1. Data ingestion and normalization: CSV files are read using Pandas, and preprocessing and normalization techniques are applied.
2. Validation functions: A set of validation rules must be established for each dataset.
3. Create error handling options: In this tool, users can choose between three ways to deal with invalid lines, named, automatic correction using synthetic data, manual correction via interactive input, and ignoring invalid lines [6].
4. Output and Reporting: The validated and corrected rows are saved to a new CSV file. A data quality report is printed summarizing the valid/invalid rows.
5. Validation Dataset: Tool responses were evaluated for accuracy, flexibility, and usability [5].

The core script is built in Python and uses modules like pandas, IP address, and re libraries. Upon uploading a csv file, the tool automatically analyzes rows and classifies them as valid or invalid based on predefined criteria. The user then selects one of three correction modes. Cleaned data is saved locally, and a report is generated summarizing the process [1].

This implementation supports common real-world use cases such as cleaning survey data, user sign-up logs, and public datasets with missing fields.

## 4 Results

The data cleaning tool developed in this project handles the missing values with synthetic data generation, which was adopted as the primary method for handling missing and invalid entries. This decision was based on its effectiveness in balancing automation and data integrity. The approach allows for the seamless replacement of missing or erroneous fields (e.g., names, email addresses, IP addresses) while preserving the overall structure and usability of the dataset. Additionally, it aligns well with privacy-conscious data practices, making it suitable for real-world scenarios where real user data is restricted [1, 3].

As an initial step, the dataset used for testing was a csv file, containing 1000 rows and 5 columns: "ID", "First/Last Name", "Gender", "Email", and "IP Address". This dataset has missing and invalid values in every row. Table 1 presents a small sample of the original dataset, illustrating rows with missing or incorrect data (e.g., null values, invalid emails, or malformed IP addresses).

**Table 1.** Small sample of the dataset.

| ID | First Name | Last Name | Email | Gender | IP Address |
|---|---|---|---|---|---|
| 2nd Floor | Fayth | Ballintyne | fballintyne0@fotki.com | Female | 159.10.192.104 |
| Suite 90 | Tony | Pummery | tpummery1@facebook.com | Male | 92.53.7.58 |
| PO Box 255 | Leigh | Elfe | lelfe2@addthis.com | Male | 205.243.222.59 |
| 13th Floor | Brendin | Leathem | bleathem3@arizona.edu | Male | |
| 3rd Floor | Natasha | Goldsbrou | ngoldsbrough4@ask.com | Female | |
| Room 1990 | Diarmid | Ferrieroi | dferrieroi5@miitbeian.gov.cn | Male | 8.226.220.110 |
| Apt 1460 | Kati | Oleshunin | koleshunin6@auda.org.au | Female | 199.98.223.148 |
| 19th Floor | Philippine | Camies | pcamies7@lulu.com | Female | |

The system checks each information in each column present in the dataset, a small sample in Table 1, taking into account:

– ID, ensures it's an integer.
– First/Last Name, non-empty strings.
– Gender, must be 'Male' or 'Female'.
– Email, checked using regex pattern.
– IP Address, validated using Python's IP address module [3].

After loading the file into the system, the system runs an automatic quality check. The output indicated that all 1000 rows were invalid, due to missing or incorrect data. This is confirmed by the report shown in Fig. 1.

```
Enter the full path of the CSV file: C:\Users\Utilizador\Desktop\MOCK_DATA.csv
Loaded CSV in 0.0038 seconds
Validation completed in 0.0460 seconds

--- DATA QUALITY REPORT ---
Total Rows: 1000
Valid Rows: 0 (0.0%)
Invalid Rows: 1000 (100.0%)
Rows with Missing or Invalid IDs: 1000

What do you want to do?
1. Auto-generate missing values using synthetic data
2. Manually fix invalid rows
3. Skip all invalid rows and keep only valid data
4. Exit
```

**Fig. 1.** Report given by the system about the data quality.

The users can select any option from the report, present in Fig. 1. Selecting option 1, automatically filled in the missing fields using predefined rules and saved the cleaned dataset to a new csv file. A portion of the cleaned file is shown in Fig. 2, with all rows corrected and validated.

As seen in Fig. 2, the system exhibits efficient runtime behavior throughout the validation and correction pipeline. The csv file was loaded in approximately 0.0038 seconds, and validation was completed in 0.0460 seconds. After selecting the option to auto-generate missing values, the correction process was completed in 2.8251 seconds, bringing the total script execution time to just 2.8784 seconds. These timings demonstrate the tool's ability to process data swiftly and accurately, even with high volumes of invalid entries.

## 5 Evaluation and Discussion

The accuracy of the synthetic data generation was evaluated by comparing the generated values with the original dataset, where available. Due to limitations in the dataset, a complete accuracy assessment was not feasible; however, initial tests indicate that the tool produces reasonable approximations of missing values, supporting its effectiveness in data cleaning tasks.

The tool offers flexibility by providing users with multiple options to handle missing or invalid data, including automatic generation of synthetic data, manual correction of entries, or skipping invalid rows. This user-driven design enables individuals to customize the cleaning process according to their specific data quality needs and project requirements.

Regarding usability and visibility, the system presents clear and concise reports detailing the status of the dataset, including counts of valid, invalid, and missing data.

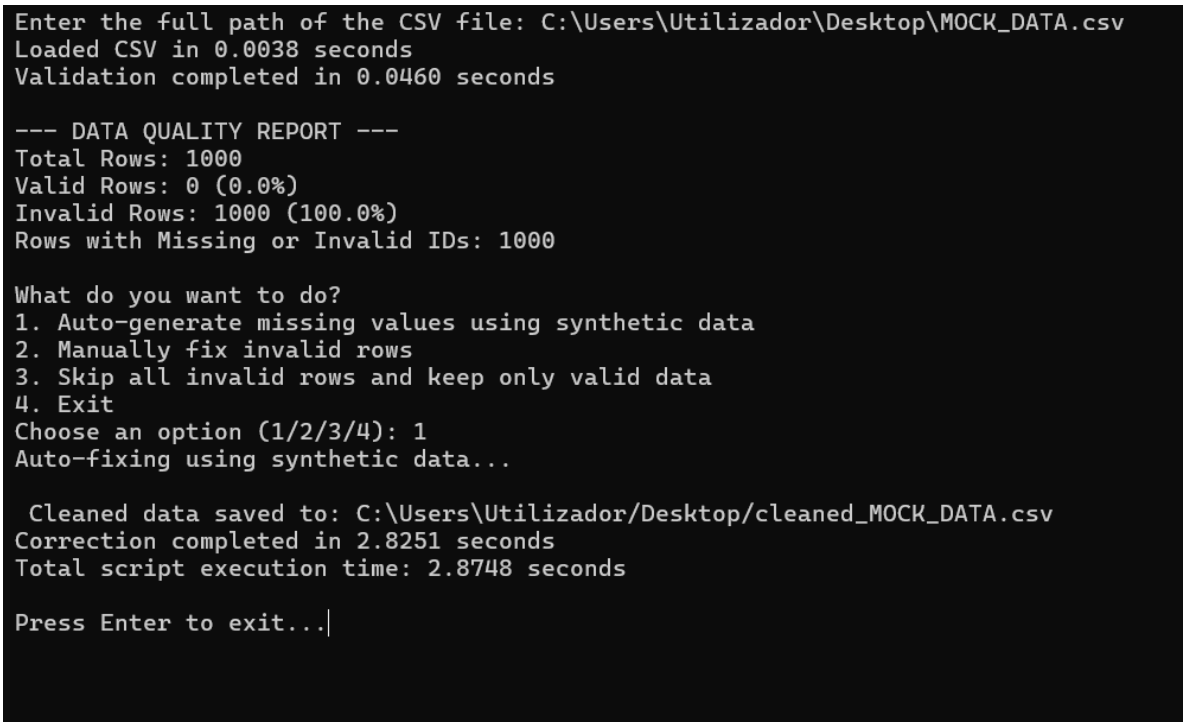



**Fig. 2.** Final report given by the system about the data quality.

User prompts guide the decision-making process, enhancing overall clarity and ease of use. Future improvements could include advanced visualization features to better represent data quality metrics and facilitate quicker insights.

At this stage, a direct comparison with other tools was not conducted due to a lack of access to detailed benchmarks, but the tool was tested with a dataset of 1000 rows and showed consistent and reliable behavior in correcting, reporting, and saving data. However, this tool focuses on quick, script-based workflows for educational or prototyping contexts, while other tools like OpenRefine and DataCleaner offer advanced GUI-based data cleaning; they often require manual configuration and lack direct support for synthetic data generation.

## 6 Conclusion and Future Works

Clean and accurate data is foundational to any successful analysis or model. This project demonstrates a practical, user-interactive tool for addressing data quality problems. By combining validation logic with flexible correction strategies, the tool prepares datasets for confident analysis, even in input errors or missing values [6].

This tool is innovative because of the integration of multiple validation and correction techniques within a single, unified system. The tool not only detects missing or invalid data, it also provides options to manually correct it, auto-generate synthetic data, or ignore it — all in a single pipeline. It ensures type-matching between headers and data (e.g., integer ID fields), checks for spelling or formatting issues, and gives the user a

clear summary with the option to update accordingly. These combined features allow for flexible and context-aware data cleaning. Additionally, it runs efficiently and provides real-time feedback, with most evaluations completing in seconds.

Future work will focus on conducting comprehensive accuracy evaluations using benchmark datasets and extending the tool's capabilities to support complex data types such as Excel spreadsheets and JSON files. Additionally, collaborative features will be incorporated to enable multiple users to work simultaneously on data cleaning tasks in a shared environment, improving efficiency and scalability.

## Acknowledgement

This work was supported by national funds: UID/05757 - Research Centre in Digitalization and Intelligent Robotics (CeDRI); and LA/P/0007/2020 (DOI: 10.54499/LA/P/
0007/2020).

## References

1. Ghosh, R.: Dealing with missing values in python: A complete guide (may 2021), https://www.analyticsvidhya.com/blog/2021/05/dealing-with-missing-values-in-python-a-complete-guide/, acesso em: 28 April 2025
2. Han, J., Pei, J., Tong, H.: Data mining: concepts and techniques. Morgan kaufmann (2022)
3. IBM: Data cleaning, https://www.ibm.com/think/topics/data-cleaning, acesso em: 28 April 2025
4. Mirzaei, A., Carter, S.R., Patanwala, A.E., Schneider, C.R.: Missing data in surveys: Key concepts, approaches, and applications. Research in Social and Administrative Pharmacy **18**(2), 2308–2316 (2022)
5. Toutenburg, H.: Rubin, db: Multiple imputation for nonresponse in surveys: Wiley, new york 1987. xxix+ 258 pp (1990)
6. Van Buuren, S., Groothuis-Oudshoorn, K.: mice: Multivariate imputation by chained equations in r. Journal of statistical software **45**, 1–67 (2011)

# Tchumy: Assistive Wearable Medical Technology for Children with Autism

Mariam Jvarsheishvili[1], Ahmed Gamal Ibrahim[1], Rui Pedro Lopes[1]

Research Center in Digitalization and Intelligent Robotics (CeDRI), Laboratório Associado para a Sustentabilidade e Tecnologia em Regiões de Montanha (SusTEC), Instituto Politécnico de Bragança, Campus de Santa Apolónia, 5300-253 Bragança, Portugal
a61615@alunos.ipb.pt, ahmed@ipb.pt, rlopes@ipb.pt

**Abstract.** Children with Autism Spectrum Disorder (ASD) face several difficulties, including sensory sensitivities and emotional regulation issues, which require specialized assistive technologies. This paper presents Tchumy, a wearable brooch designed to monitor environmental sound and light levels in real time, to help caregivers prevent sensory overload. The device is built around the ESP32 microcontroller and is supported by a mobile app. Tchumy is a customizable, non-intrusive solution, fitting to each child's needs, to help them navigate through life more independently.

**Keywords:** Autism Spectrum Disorder · Pediatric Medical Devices · Assistive Technology · Wearables · Tchumy

## 1 Introduction

Medical devices have evolved throughout the centuries, but pediatric innovations often develop more slowly, because of challenges related to ethics, technology, and market limitations. Children, specifically with ASD, need specialized, comfortable, and adaptive assistive technologies to manage sensory sensitivities. ASD is characterized by heightened responses to environmental stimuli, such as sound and light, which can lead to distress. Wearable technologies offer a discreet method to monitor these stimuli and timely alert the caregiver for support. This paper introduces Tchumy, a customizable, brooch-style wearable device, designed for children with ASD, that tracks environmental sound and light levels in real time and provides alerts through a mobile app, to improve children's emotional well-being, environmental awareness and independence. The structure of this paper is as follows: Section 2 reviews the state-of-the-art in medical devices, specifically the ones used among children. Section 3 discusses the methodology and the development. Section 4 summarizes key points and insights and outlines future research directions.

## 2 State of the Art

Despite centuries old usage and development of medical devices, pediatric devices (medical devices intended for use in persons aged 21 years or younger at the time of their diagnosis or treatment) face unique challenges across the total product life cycle of device innovation and development [6]. Medical devices for pediatric populations must account for children's anatomical, physiological, and psychological differences compared to adults [6]. Pediatric devices require specific, specialized design: smaller sizes, higher

sensitivity, minimal invasiveness, and user-centered comfort. Examples include infant blood pressure cuffs, pediatric otoscope tips, and modified endoscopes [6]. Although they are important, pediatric medical devices fall behind adult-focused innovations, due to limited market size, increased design challenges, and ethical regulations for clinical trials involving minors [6].

Medical devices for children, especially assistive technologies, include many different innovative systems and wearables that help children with special needs with daily activities, learning, and quality of life. AT solutions range from speech-generating devices and eye-tracking systems to sensor-based tools embedded in everyday environments [1, 14]. Children's bodies are not simply smaller versions of adults, they have different proportions, physiological parameters, and growth patterns that affect device design requirements. Their design must ensure durability, ease of use, affordability, and adaptability to developmental changes [14].

These technologies help them to overcome limitations and promote independence, social integration, and development [14]. ASD is a common lifelong neurodevelopmental disorder, difficulties with social communication and interaction, a strong attachment to routines, and hyper-sensory sensitivities [16]. Sensory processing differences are central in ASD, influencing how affected individuals perceive and respond to environmental stimuli [11]. Children with ASD often become distressed or upset by slight changes in their routine, as well as extreme sensory input (e.g., noise, light, etc.) from the environment [11].

Wearable Technology (WT) offers promising, non-intrusive ways to monitor and support children with ASD. These devices capture physiological and behavioral data in real-time, aiding early detection of stress, promoting emotional regulation, and improving communication [1, 11]. Common wearable formats include smartwatches, biosensor wristbands, and smart glasses. Integrated sensors measure physiological markers such as Heart Rate (HR), Electrodermal Activity (EDA), Respiration Rate (RR), body temperature, SpO2, and Electromyogram (EMG) [1, 11]. These data are processed through Machine Learning (ML), Internet of Things (IoT), and cloud computing to inform caregivers and therapists in real-time [14]. Device materials and ergonomics are critical for ASD populations, many of whom are hypersensitive to texture or pressure, therefore, if not chosen correctly, the device that is supposed to help children, may be an additional trigger. Wrist-worn or fabric-integrated devices are generally better tolerated than chest or head-mounted alternatives [11].

Designing wearable technologies for pediatric healthcare, especially for children with ASD, needs a lot of thought and research regarding the hardware, software, and communication systems, which should be suitable for continuous and non-invasive monitoring. These wearables collect real-time data on different factors, to identify stressors or unusual patterns, which will prompt caregivers to intervene. To do this, wearable devices need microcontrollers that have to perform energy efficiently. The ESP32 family, developed by Espressif Systems, is a popular choice due to its dual-core processing, low-power capabilities, and built-in support for Bluetooth Low Energy (BLE) and Wi-Fi [7]. These features allow wearables to handle multiple data streams from sensors while maintaining long battery life, which is critical for continuous

monitoring in pediatric applications. For instance, the ESP32 can preprocess sensor data locally, which will reduce the need to transmit data all the time. This not only conserves energy, but also minimizes privacy risks by limiting the amount of raw data sent to external servers [8].

Wearables for children with ASD rely on a plethora of sensors to monitor the environment and the child's physiological state. Ambient sensors, such as sound and light sensors, are important for detecting overstimulating conditions that may trigger distress in children with sensory sensitivities [1]. The sensors should monitor passively, so the device will integrate into a child's daily life without causing disruption or additional triggers. The data that the devices generate, needs to be processed and presented meaningfully through companion applications.

The companion applications are just as important as the hardware. The softwares, often developed using cross-platform frameworks like React Native, allow developers to create apps for both iOS and Android with a single codebase [12]. This approach ensures a consistent user experience across platforms and simplifies updates. These apps often integrate with cloud platforms like Firebase for secure data storage and real-time synchronization with dashboards used by caregivers or clinicians [9]. For example, Firebase's real-time database and authentication features enable secure, role-based access to sensitive health data. The choice of software framework also influences how data is transmitted from the wearable to the cloud.

To transmit data to companion apps or cloud services, BLE and Wi-Fi are the most common options. BLE, supported by the ESP32, has low energy consumption, which is ideal for short-range, real-time data exchange [3]. But BLE requires a paired mobile device to relay data to the cloud. On the other hand, Wi-Fi offers higher bandwidth and direct internet access, but it has high power consumption, which makes it less suitable for longer use without frequent recharging. For prototyping or less power-sensitive applications, HTTP over Wi-Fi may be used for its compatibility with RESTful APIs, which simplifies integration with cloud services like Firebase. These choices, however, must also take into account platform-specific challenges, particularly on iOS.

Developing wearables for iOS has a lot of obstacles, due to Apple's restrictions on background activity and hardware access. For instance, iOS limits continuous BLE operations, which can disrupt real-time data streaming [2]. Developers often adopt hybrid approaches during prototyping, prioritizing development speed over production-ready efficiency [11]. These obstacles show the need for careful planning in communication design, which must also address security and privacy to protect sensitive data. Given the sensitivity of health data, security is very important in wearable technologies design. Secure protocols like HTTPS and TLS are standard for data transmission, while cloud platforms like Firebase provide built-in authentication and access control [10]. BLE version 4.2 and later offers enhanced encryption and secure pairing, and Wi-Fi connections use WPA3 for security [15]. Compliance with regulations like the Children's Online Privacy Protection Act (Children's Online Privacy Protection Act (COPPA)) and the Health Insurance Portability and Accountability Act (Health Insurance Portability and Accountability Act (HIPAA)) is

non-negotiable, ensuring that data handling meets legal and ethical standards [8]. These security measures tie back to the overall goal of creating trustworthy, effective systems for pediatric care.

## 3 Methodology and Development

Sensory hypersensitivity is a symptom often associated with ASD, as more than 90% of children who have ASD experience this sensory challenge, which can significantly impact their daily lives [13]. The most common sensory sensitivities are related to sound and light. For instance, children may have sensitivity to everyday sounds, such as the hum of fluorescent lights or loud conversations. This heightened sensitivity can cause distress, emotional outbursts, and anxiety. Similarly, bright or flickering lights can overwhelm children with ASD, leading to sensory overload and potentially triggering meltdowns. These challenges can severely limit their ability to participate in typical social and school environments.

The idea for the development of Tchumy, a wearable device for children with ASD, was to create a supportive and non-invasive device, that would integrate into a child's daily life and monitor their environment for light and sound levels around them. Tchumy is a brooch, worn on clothing, with a 3D-printed casing, which will allow children to showcase their interests. Tchumy's design prioritizes the unique needs of children with ASD, especially their sensory sensitivities. Material or placement of the wearable can be an additional trigger for the child, therefore it was necessary to think of something non-invasive. A brooch or pin format was chosen over wristbands or head-mounted devices, as it avoids direct skin contact and can be pinned to clothing in a position comfortable for the child. The casing is made from lightweight materials via 3D printing, which allows customization with designs like favorite characters or patterns, that will create a sense of ownership for the children, therefore reducing the resistance towards wearing the device. An important design consideration is whether the system should interact with the child when elevated light or sound stimuli are detected. Currently, Tchumy operates passively, transmitting data without providing feedback to the child. While this avoids potential overstimulation or distraction, such additions may be able to provide a calming experience to the child, if tailored to their needs. However, the interactions must be carefully evaluated to avoid introducing unintended stressors.

Tchumy uses hardware components, which were selected for their performance, power efficiency, and suitability for wearable devices. The device works with the ESP32 microcontroller, which has processing capability and low-power consumption. The main reason for using this microcontroller is its functionality and small size, it has built-in Wi-Fi and Bluetooth connectivity. The ESP32 can preprocess the sensor data locally, reducing cloud transmission, therefore extending not only the battery life, but also minimizing the privacy issues, by limiting the raw data exposure. For environmental monitoring, Tchumy uses a microphone sound sensor module (VMA309) and an Ambient Light Sensor (BH1750) to detect stimuli that may trigger sensory sensitivities in children with ASD. The sound sensor catches ambient noise levels across different frequencies, giving us detailed information, not only simple volume measurements. Similarly, the light sensor measures intensity and spectral

characteristics of lighting, enabling the detection of potentially problematic light sources, such as fluorescent lighting or rapid flashing [5].

Tchumy samples sound data at a rate of 5-10 Hz, collecting 5 to 10 readings per second. This sampling rate captures sufficient data to detect patterns and also maintains system efficiency on the ESP32 microcontroller. The system processes each reading in real-time, calculating averages and derivatives to differentiate between steady, repetitive sounds (which may indicate problematic environmental stimuli) and irregular, transient ones like loud television noise. The last 20 seconds of sound data are stored in a rolling buffer.

When the amplitude of the sound exceeds a set threshold and is identified as potentially distressing, the ESP32 sends the buffered data via HTTP to a mobile application built with Expo Go. The project uses Firebase as the backend for storage and communication. Although BLE is commonly used for low-power, short-range communication in embedded systems, it was not utilized in this project due to Expo Go's limited BLE support, leading to the selection of Wi-Fi instead. The threshold is individually configurable by caregivers through the mobile app interface. This customization is based on the unique sensory profile of each child, information that can be informed by clinical evaluations or sound sensitivity assessments conducted by psychologists or audiologists [4].

For timely intervention, caregivers can set the alert threshold slightly below the child's known discomfort level. For example, if a child is known to react negatively to sounds at 90 dB, the threshold may be set at 85 dB to provide caregivers with a preemptive alert, which will allow them to intervene before the child becomes distressed. The ESP32 employs multicore processing to ensure that sampling, analysis, and transmission tasks run smoothly in parallel, while queues and event triggers manage efficient data handling without overwhelming the system.

The Tchumy mobile app is an important companion to the wearable brooch designed for children with ASD. It connects seamlessly with the device's ESP32 microcontroller to deliver real-time information about the child's environment, helping parents or caregivers manage sensory sensitivities effectively. The app is built to be intuitive and practical, it offers tools to monitor, analyze, and customize the device's settings to suit each child's unique needs.

The app's main screen provides a live view of ambient sound and light levels, detected by the device's VMA309 sound sensor and BH1750 light sensor. This real-time display updates continuously, showing whether the environment is within a safe range for the child or if conditions, like loud noises or bright lights, might be overwhelming. Caregivers can set custom thresholds for sound and light through the app, tailoring alerts to the child's specific sensitivities. For example, a parent might adjust the sound threshold to flag noises above a certain decibel level, ensuring the app highlights potential triggers promptly.

Beyond immediate feedback, the app offers a historical sound level chart to help caregivers spot patterns over time. This visual tool plots sound data, making it easier to identify recurring environmental triggers, such as noisy school hallways or flickering lights, that may cause distress. By reviewing these trends, caregivers can make informed

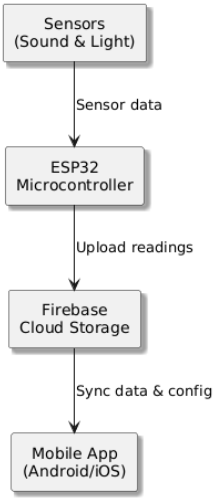



**Fig. 1.** System architecture of Tchumy.

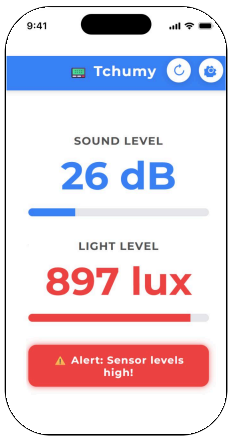



**Fig. 2.** Main screen of Tchumy app.

adjustments to the child's surroundings to reduce sensory overload. All data is securely stored in Firebase, the app's backend, which ensures reliable storage and real-time syncing between the device and the app.

The app's settings screen allows caregivers to fine-tune the device to match the child's sensory profile, with just a simple slider to adjust the thresholds. Changes made in the app are instantly sent to the Firebase, making sure the device sends back the updated settings in real time. This flexibility allows Tchumy to adapt to each child's needs, to be a supportive tool, not a one-size-fits-all solution. When sound or light levels exceed the set thresholds, the app sends push notifications to caregivers immediately. These alerts can be customized to trigger only at specific levels, such as when noise hits a certain decibel range, so the caregivers are informed without getting constant updates disguised as false alarms. A service worker delivers the notifications even if the app is running in the background, so caregivers can stay aware of potential issues without keeping the app open. To validate the system's practical usability, future phases will include structured testing involving children with ASD and their caregivers. This evaluation will focus on assessing the device's comfort, usability, and effectiveness in managing sensory sensitivities in real-world environments.

To validate the system's practical usability, future phases will include structured testing involving children with ASD and their caregivers. This evaluation will focus on assessing the device's comfort, usability, and effectiveness in managing sensory sensitivities in real-world environments.

## 4 Conclusions and Future Work

Tchumy's development presents how necessary it is to design pediatric medical devices that are not only functional, but also sensitive to the specific needs of children with ASD. As a supportive tool, Tchumy aims to assist caregivers and parents in managing daily experiences for children with ASD by minimizing sensory overload

through real-time environmental monitoring. Its key features: a non-invasive brooch design, customizable thresholds, and an intuitive mobile application, which are made with principles of child-centered care.

While this work presents the methodology and design of Tchumy, further validation is required to assess its effectiveness. In future studies, we will implement a structured evaluation strategy, including user testing with children with ASD, their caregivers and clinical professionals, to measure usability, comfort, and real-world performance. Additionally, we will transition from Wi-Fi to BLE for improved energy efficiency and smooth connectivity.

Looking ahead, Tchumy's development will focus on embedding data analytics for more precise detection and intervention. These advancements will refine the device's functionality, enhance its software, and strengthen its role in personalized autism support. By prioritizing inclusive and accessible design, Tchumy demonstrates how wearable technology can evolve to better serve children with sensory processing challenges.

## Acknowledgement

The authors are grateful to the Foundation for Science and Technology (FCT, Portugal) for financial support through national funds FCT/MCTES (PIDDAC) to CeDRI, UIDB/05757/2020 (DOI: 10.54499/UIDB/05757/2020) and UIDP/05757/2020 (DOI: 10.54499/UIDP/05757/2020) and SusTEC, LA/P/0007/2020 (DOI: 10.54499/LA/P/0007/2020).

## References

1. Ahuja, A., Krishnan, N., Wang, Y., Jain, V., Moturu, S.P., Nirala, V., Subramanian, V., Sundaram, H., Subramanian, L.: Wearable devices for children with autism spectrum disorder (asd): A literature review. ACM Transactions on Computing for Healthcare (HEALTH) **3**(4), 1–33 (2022). https://doi.org/10.1145/3510426
2. Apple Inc.: Core Bluetooth Framework Reference. Apple Inc. (2025), `https://developer.apple.com/documentation/corebluetooth`, accessed: 2025-05-14
3. Bluetooth Special Interest Group (SIG): Bluetooth Core Specification Version 6.0. Bluetooth SIG, Inc. (August 2024), `https://www.bluetooth.com/specifications/specs/core-specification-6-0/`, accessed: 2025-05-14
4. Chaldi, D., Mourtzouchos, K., Lygeros, S., Danielides, G., Naxakis, S.: Evaluation of Hearing Thresholds in Infants With Autism Spectrum Disorder Using Auditory Brainstem and Steady-State Responses. Cureus **17**(1), e77537. https://doi.org/10.7759/cureus.77537, `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11829610/`
5. Deng, L., Rattadilok, P.: A Sensor and Machine Learning-Based Sensory Management Recommendation System for Children with Autism Spectrum Disorders. Sensors **22**(15), 5803 (Jan 2022). https://doi.org/10.3390/s22155803, `https://www.mdpi.com/1424-8220/22/15/5803`, number: 15 Publisher: Multidisciplinary Digital Publishing Institute
6. Duffy, S., Krishnan, A., Yazdi, Y., Quan, X., Hughes, M., Marsal, A.L., Peiris, V., Frassica, J.J., Eskandanian, K., Sen, D.G.: The challenges and opportunities in pediatric medical device innovation: Monitoring devices. The Annals of Thoracic Surgery (2024). https://doi.org/10.1016/j.athoracsur.2024.11.034, `https://linkinghub.elsevier.com/retrieve/pii/S0003497524011056`
7. Espressif Systems: ESP32 Technical Reference Manual. Espressif Systems (2025), `https://www.espressif.com/sites/default/files/documentation/esp32_technical_reference_manual_en.pdf`, accessed: 2025-05-14

8. Gerodimos, A., Maglaras, L., Kantzavelou, I., Ayres, N.: Iot: Communication protocols and security threats (01 2022). https://doi.org/10.20944/preprints202111.0214.v2
9. Google LLC: Firebase Documentation. Google LLC (2025), `https://firebase.google.com/docs`, accessed: 2025-05-14
10. Google LLC: Firebase Security Rules Documentation. Google LLC (2025), `https://firebase.google.com/docs/rules`, accessed: 2025-05-14
11. Koumpouros, Y., Kouroupetroglou, C., Kyttari, E.: Wearable systems for monitoring mobility-related activities in children with autism spectrum disorder: A systematic review. Journal of Healthcare Engineering **2019**, 1–17 (2019). https://doi.org/10.1155/2019/2356702
12. Meta Platforms, Inc.: Getting Started with React Native. Meta Platforms, Inc. (2025), `https://reactnative.dev/docs/getting-started`, accessed: 2025-05-14
13. National Institute of Mental Health Press Office: Understanding the Underpinnings of Sensory Hypersensitivity in SCN2A-Associated Autism (Apr 2024), `https://shorturl.at/M0XwS`
14. Pandey, R., Mishra, P., Verma, R., Jain, S.: Wearable and assistive devices for children with special needs: An overview. International Journal of Pediatrics and Adolescent Medicine **11**(1), 12–25 (2024). https://doi.org/10.1016/j.ijpam.2023.09.003
15. Wi-Fi Alliance: Wi-Fi Security Overview. Wi-Fi Alliance (2025), `https://www.wi-fi.org/discover-wi-fi/security`, accessed: 2025-05-14
16. World Health Organization: Autism spectrum disorders (2025), `https://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders`, accessed: May 13, 2025

# Exploratory Data Analysis and Insights on Volatile Organic Compounds for Hazardous Waste Detection

Mahdia Ahmadi[1], Natalia Méndez Pérez[1,2], Helena Cristina Almeida da Cruz[3], Getúlio Igrejas[1], Pedro João Rodrigues[1], and Rui Pedro Lopes[1]

[1] Research Centre in Digitalization and Intelligent Robotics (CeDRI), Laboratório Associado para a Sustentabilidade e Tecnologia em Regiões de Montanha (SusTEC), Instituto Politécnico de Bragança (IPB), Bragança, Portugal
{mahdia, igrejas, pjsr, rlopes}@ipb.pt
[2] Univerdidad de Lalaguna, Tenerife, Spain
lu0101487038@ull.edu.es
[3] Instituto Politécnico de Coimbra, Coimbra, Portugal
cac.helena12@gmail.com

**Abstract.** Volatile Organic Compounds (VOCs) are critical indicators of environmental contamination, particularly in hazardous waste contexts. While gas chromatography–mass spectrometry (GC–MS) provides high specificity, it struggles with scalability and pattern discovery in large, complex datasets. This study presents a data-driven framework integrating Exploratory Data Analysis (EDA) techniques — including principal component analysis (PCA), hierarchical clustering, and correlation mapping — to uncover emission patterns in compost-derived VOC data.

Using the LCSC VOC 2022 Compost Dataset (141 variables, 90 samples), we identified strong co-emission clusters (e.g., D-Limonene and $\alpha$-Pinene) and a temperature-dependent ethanol emission pattern unique to food-and-yard waste samples. Pearson correlation analysis revealed shared emission behavior, and regression confirmed a positive slope (25.6) for ethanol versus temperature.

These findings highlight EDA's potential to enhance VOC dataset interpretability and source identification. The proposed framework supports practical applications such as early-warning systems, sensor deployment, and data-informed environmental policy.

**Keywords:** Hazardous Waste, Pollution Source Identification, Data-Driven Decision Making, Machine Learning, Sensors.

## 1 Introduction

Volatile Organic Compounds (VOCs) are a broad class of carbon-based chemicals that readily vaporize at ambient temperatures. They are emitted from a wide range of natural and anthropogenic sources, including industrial activities, landfills, and composting systems. The presence of VOCs in the environment is of increasing concern due to their adverse impacts on both ecological systems and human health. Studies have established links between long-term exposure to VOCs and serious health conditions such as chronic respiratory illnesses, skin cancer, and neurological disorders [6].

To monitor VOCs, gas chromatography-mass spectrometry (GC–MS) has long been considered the gold standard, offering high specificity and sensitivity. However, the growing complexity and dimensionality of environmental data pose serious challenges for traditional techniques. Specifically, GC-MS struggles with scaling to

large datasets, lacks real-time analytical capacity, and is limited in its ability to differentiate between complex emission sources in mixed environments [1]. These shortcomings are particularly evident in hazardous waste contexts, where data is often heterogeneous, poorly labeled, and influenced by dynamic environmental variables.

In response to these limitations, data-driven analytical frameworks are gaining traction in environmental monitoring. Among these, Exploratory Data Analysis (EDA) offers a compelling approach. EDA is a flexible, assumption-free method that leverages statistical visualization and unsupervised learning to reveal underlying structures, correlations, and outliers within datasets [5]. Its non-parametric nature makes it especially useful for VOC analysis, where datasets are multidimensional, noisy, and lack uniform labels. In biomedical and environmental contexts, EDA has shown promise in applications ranging from air quality modeling to disease biomarker detection via VOCs [10].

In this work, we apply EDA techniques — including principal component analysis (PCA), hierarchical clustering, and correlation mapping — to a real-world VOC dataset obtained from composting systems. Unlike prior studies that focus solely on compound identification or source quantification, our approach integrates chemical analysis with pattern discovery to uncover emission trends, temperature-driven behavior, and co-emission clusters. Our main contribution lies in demonstrating how EDA can supplement traditional chemical monitoring, enabling scalable, interpretable, and actionable insights that are critical for hazardous waste management.

The rest of this paper is structured as follows. Section 2 discusses relevant background literature and related work. Section 3 outlines the methodological framework and dataset selection process. Section 4 presents key findings from the data analysis, while Section 5 concludes the study with insights and future research directions.

## 2 Background and related work

Volatile Organic Compounds (VOCs) are a well-established environmental and public health concern due to their role in air pollution and potential toxic effects on humans. VOCs originate from various anthropogenic sources such as petroleum refineries, landfills, and industrial waste processes, as well as from natural biological emissions [3] [10].

In recent years, there has been growing attention on the health implications of chronic VOC exposure. For instance, large-scale studies have revealed associations between long-term VOC exposure and increased risks of respiratory diseases and skin cancer, especially in vulnerable populations. These findings underscore the importance of precise and scalable VOC detection techniques, particularly in sensitive environments like hazardous waste sites.

Conventional detection methods such as gas chromatography-mass spectrometry (GC-MS) offer high analytical specificity, but they are resource-intensive, slow, and poorly suited for high-throughput or real-time applications. More importantly, they fall short when interpreting complex, heterogeneous, and multidimensional datasets,

which are often the norm in environmental monitoring contexts involving mixed waste types, variable emission profiles, and fluctuating environmental conditions [1].

To address these limitations, data-driven methods, including machine learning and statistical modeling, have been increasingly explored. For example, VOC-based classification has been used for food quality inspection [7], tracing coffee origin through network analysis [8], and disease detection via exhaled breath [5]. However, these methods often require labeled data, clear training objectives, or domain-specific features, which may not always be available in hazardous waste settings.

This creates an important gap — the limited use of Exploratory Data Analysis (EDA) in VOC research, especially in unstructured or poorly labeled datasets. Unlike predictive models, EDA does not assume prior distributions or require labels, making it well-suited for uncovering hidden structures, correlations, and co-emission patterns in environmental VOC data. Additionally, EDA allows for early insight extraction and hypothesis generation without committing to predefined models, an advantage when dealing with novel or dynamic emission environments like composting systems and hazardous waste facilities.

The main challenges include: **data heterogeneity** (varying VOC profiles by source and interaction), **sparse labeling** (limited annotations on emission origins), **environmental noise** (temperature and humidity confound analysis), and **high dimensionality** (numerous simultaneous variables). Despite these challenges, few studies have applied EDA as a central methodology in VOC analysis. This paper aims to fill that gap by demonstrating how EDA — through methods such as PCA, hierarchical clustering, and correlation analysis — can generate interpretable insights into VOC behavior, especially in compost-derived air samples. Our contribution lies in showing that EDA not only reveals complex compound interdependencies, but also aids in identifying source-specific emissions and conditions that affect their distribution.

## 3 Methodology

This study adopted an Exploratory Data Analysis (EDA) approach with the primary objective of examining and understanding the structure and content of multiple datasets containing information on VOC emissions.

EDA techniques help to provide an understanding of data, without requiring the application of formal statistical procedures or the prior definition of assumptions about the data at hand. This methodology involves the use of graphical and non-graphical methods, such as descriptive statistics, to facilitate a comprehensive understanding of the data structure, quality, and relationship between its variables, among others. Therefore, it serves as a crucial initial step in the data analysis process [5].

At the initial stage of this research, three datasets were explored: the LCSC VOC Compost Dataset 2022 [4], Long-term variations of ambient VOCs [9], and Experiments on VOC uptake by the active layer soils of Greenlandic permafrost areas [2]. However, only the first was selected for analysis. The excluded datasets were not directly relevant to the study's objectives, one focused on urban air trends, and the other on VOC absorption in Arctic soils, neither directly related to the study's focus on VOC emissions

from hazardous waste sources. In contrast, the chosen dataset offers detailed, time-resolved VOC measurements from composting activities in a hazardous waste context, making it the most suitable for identifying emission patterns, clustering behavior, and contamination sources.

## 4 Dataset Analysis

The dataset analyzed in this study, the LCSC VOC 2022 Compost Dataset, was developed by the Lewis-Clark State College (LCSC) Air Research Group, led by Dr. Nancy A. C. Johnston. It is part of a broader NIH-funded project supported by the Institutional Development Award (IDeA) from the National Institute of General Medical Sciences in partnership with LCSC. The data was collected at the Washington State University Compost Facility to investigate the VOCs emitted from compost under different conditions.

Data collection occurred from July to September 2022, using high-resolution sampling intervals. The dataset comprises 90 samples: 84 combined air and water samples, 4 air-only samples, and 2 water-only samples. In total, 141 variables were measured, including chemical, environmental, and sampling-related parameters.

VOC concentrations are reported in parts per billion by volume (ppbv), with a measurement uncertainty of ±10%. The VOC concentrations in air samples were measured using thermal desorption tubes with a Markes-Agilent TD-GC-MS system, while water-phase VOCs were captured using impinger sampling and analyzed with an Agilent HS-FID-GC system.

This study primarily focuses on the analysis of air samples, as they are most relevant to our ongoing and future research. A total of 88 air samples were collected.

### 4.1 Sample Conditions Analysis

Table 1 presents the descriptive statistics of the air sample variables, while Fig. 1 illustrates the origin of the compost samples and their specific locations within the pile.

**Table 1.** Descriptive statistics for environmental variables in air samples

| | **Pile Temp. [ºC]** | **Outside Temp. [ºC]** | **Humidity [%]** | **Pressure [atm]** | **Wind Speed [m/s]** |
|---|---|---|---|---|---|
| **mean** | 50.35 | 28.22 | 26.32 | 0.91 | 3.98 |
| **std** | 17.60 | 4.35 | 10.75 | 0.002 | 1.65 |
| **min** | 22.22 | 21.70 | 10 | 0.90 | 0.45 |
| **25%** | 34.72 | 22.68 | 20 | 0.90 | 3.13 |
| **50%** | 55 | 28.75 | 20 | 0.91 | 4.02 |
| **75%** | 65.56 | 31.63 | 36.25 | 0.91 | 4.92 |
| **max** | 88.33 | 38 | 49 | 0.91 | 7.15 |

The average pile temperature (around 50°C) aligns with expected thermophilic composting conditions, while outside temperature, humidity, pressure and wind speed values fall within typical ranges for outdoor composting in warm environments.

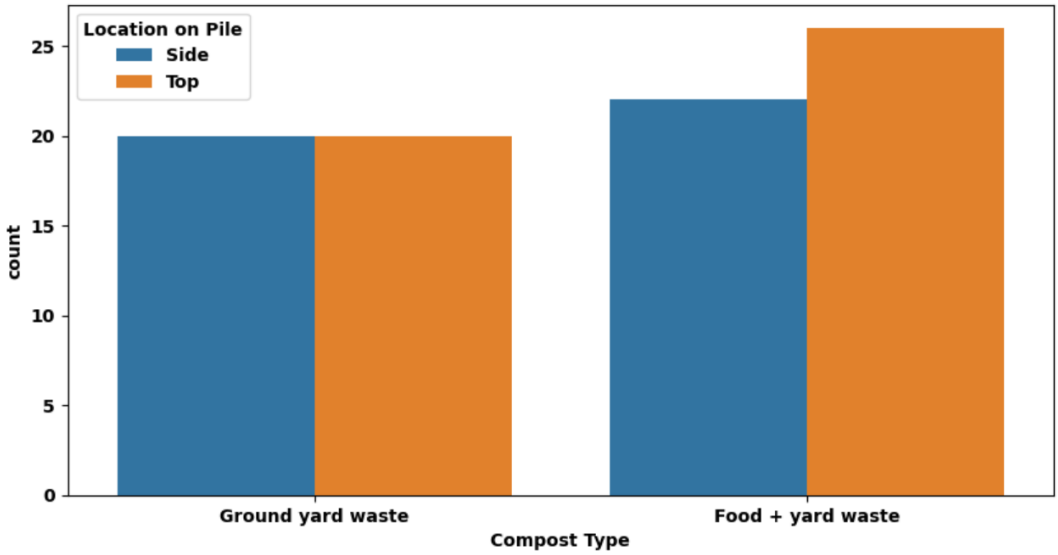



**Fig. 1.** Air sample frequency by compost type and pile location

Approximately half of the samples were taken from ground yard waste, while the other half corresponded to a mix of food and yard waste. Within each group, the samples are evenly distributed between the top and the side of the pile, ensuring a balanced representation of sampling locations across waste types.

### 4.2 VOCs concentration

Fig. 2 presents the top ten VOCs based on their average concentration (ppbv) in the samples. Fig. 3, on the other hand, illustrates the Pearson correlation coefficients observed between these VOCs.

Strong correlations are observed between the following VOC pairs: D-Limonene and $\alpha$-Pinene; $\alpha$-Pinene and $\beta$-Pinene; $\beta$-Pinene and $\gamma$-Terpinene; $\beta$-Pinene and Sabinene; $\gamma$-Terpinene and Sabinene; Camphor and L-Fenchone; Camphor and $\alpha$-Humulene; and $\alpha$-Humulene and L-Fenchone.

### 4.3 Specific Analysis for Ethanol

Fig. 4 plots ethanol concentration (the VOC with the highest average concentration) against the pile temperature, with data distinguished by waste type. A regression line is also shown to illustrate the overall trend for each compost type.

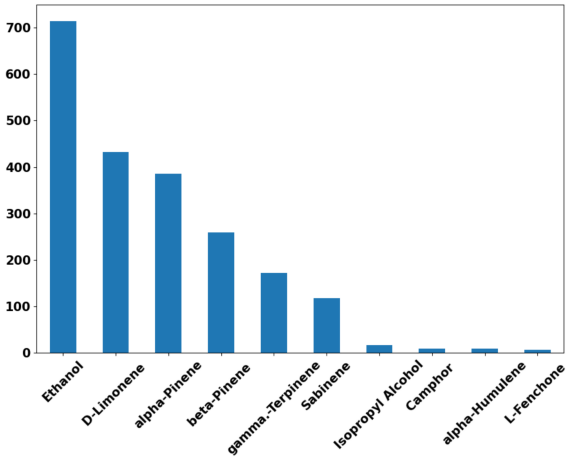



**Fig. 2.** Top ten VOCs by average concentration

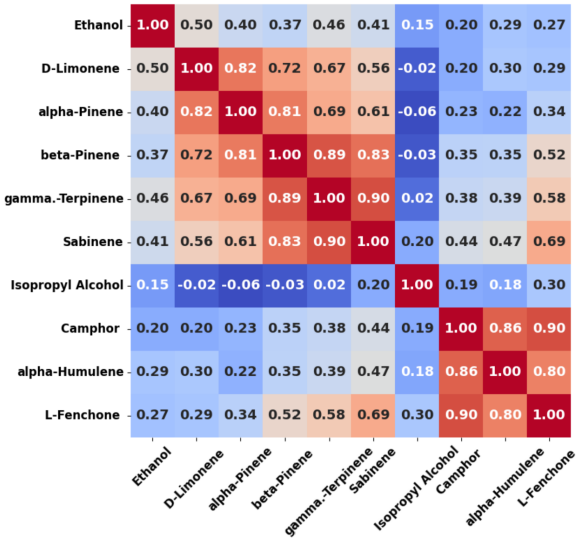



**Fig. 3.** Correlation between top 10 VOCs

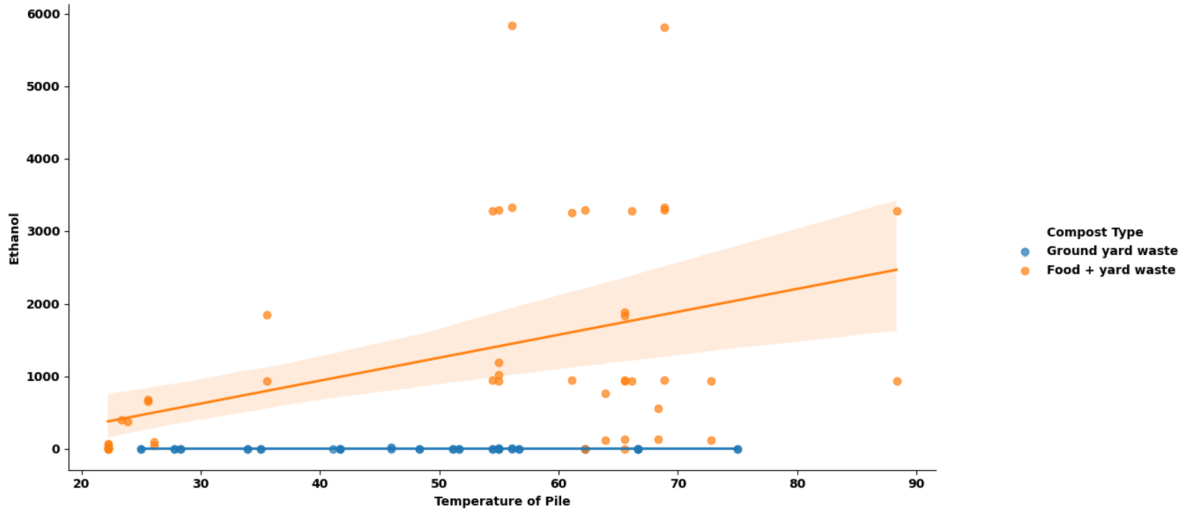



**Fig. 4.** Relationship between ethanol concentration and pile temperature by waste type, with fitted regression lines

Ethanol concentration shows a slight positive correlation with temperature, with a regression slope of 25.6292. Additionally, ethanol was not detected in ground waste samples, indicating no emissions from this type.

## 5 Results and Findings

The exploratory data analysis provided several noteworthy insights into the behavior of volatile organic compounds (VOCs) within compost environments. By focusing on a carefully selected dataset — comprising air samples collected from

compost piles of varying composition and sampling locations — key environmental patterns and compound relationships were uncovered.

Descriptive statistics highlighted distinct differences between internal and external environmental conditions. The pile temperature exhibited a broad range, with higher variability compared to the more stable external temperature. This internal heat may influence the microbial activity responsible for VOC production. Humidity and wind speed showed moderate fluctuations, while atmospheric pressure remained nearly constant, suggesting limited relevance to VOC dynamics in this context.

The dataset revealed a diverse VOC profile. Among the top 10 compounds identified by average concentration, ethanol, D-Limonene, and $\alpha$-Pinene were particularly prominent. These compounds are commonly associated with microbial fermentation and the degradation of plant-based organic matter, indicating active biological decomposition within the compost.

A closer examination of VOC relationships through correlation analysis showed strong co-emission patterns between specific compounds, such as D-Limonene and $\alpha$-Pinene, and $\beta$-Pinene with $\gamma$-Terpinene and Sabinene. These correlations suggest shared emission sources, likely tied to the decomposition of similar organic substrates, such as citrus residues or terpene-rich plant matter.

Ethanol emerged as the compound with the highest overall concentration. Its levels were found to increase with rising pile temperatures, hinting at a temperature-dependent fermentation process. Notably, ethanol was absent in samples from ground yard waste, while present in mixed food-and-yard waste piles. This suggests that food waste is the primary contributor to ethanol emissions, and that different waste compositions may lead to distinct VOC signatures.

The dataset's design ensured balanced spatial sampling across pile sides and tops. No significant concentration differences were observed based solely on sample location, reinforcing the idea that waste type and internal conditions are stronger drivers of VOC behavior than sampling orientation.

## 6 Conclusion and Future Work

This study applied exploratory data analysis (EDA) to a compost VOC dataset, yielding meaningful insights into emission patterns. Ethanol and several terpenes emerged as dominant compounds, particularly in mixed food-and-yard waste samples. Temperature and waste composition were found to significantly influence VOC levels.

Strong correlations among specific VOCs suggested shared sources, offering potential for simplified monitoring using key indicator compounds. The findings support the value of data-driven methods in complementing traditional chemical analysis for environmental monitoring.

Future work will expand the analysis using diverse datasets across seasons, compost types, and locations, while integrating machine learning and real-time sensors to improve detection and decision-making in hazardous waste environments.

## Acknowledgment

This work has been supported by FCT - Fundação para a Ciência e Tecnologia within the Project Scope: 2024.07316.IACDC/2024.

## References

1. Capitain, C., Weller, P.: Non-targeted screening approaches for profiling of volatile organic compounds based on gas chromatography-ion mobility spectroscopy (gc-ims). Molecules **26**(18), 5457 (2021)
2. Jiao, Y., Kramshøj, M., Davie-Martin, C., Elberling, B., Rinnan, R.: Dataset: experiments on volatile organic compounds uptake by the active layer soils of greenlandic permafrost areas (Nov 2024)
3. Jindamanee, K., Keawboonchu, J., Pinthong, N., Meeyai, A., Inchai, P., Thepanondh, S.: Environmental impacts and emission profiles of volatile organic compounds from petroleum refineries. Scientific Reports **15**(1) (2025)
4. Johnston, N.: LCSC VOC Compost Dataset 2022 (2023), mendeley Data, V1
5. Komorowski, M., Marshall, D.C., Salciccioli, J.D., Crutain, Y.: Exploratory data analysis, pp. 185—203. Springer International Publishing (1 2016)
6. Nalini, M., Poustchi, H., Bhandari, D., Blount, B.C., Kenwood, B.M., Chang, C.M., Gross, A., Ellison, C., Khoshnia, M., Pourshams, A., Gail, M.H., Graubard, B.I., Dawsey, S.M., Kamangar, F., Boffetta, P., Brennan, P., Abnet, C.C., Malekzadeh, R., Freedman, N.D., Etemadi, A.: Exposure to volatile organic compounds and chronic respiratory disease mortality, a case-cohort study. Respiratory Research **26**(1) (2025)
7. Shtepliuk, I., Domènech-Gil, G., Almqvist, V., Kautto, A.H., Vågsholm, I., Boqvist, S., Eriksson, J., Puglisi, D.: Electronic nose and machine learning for modern meat inspection. Journal of Big Data **12**(1) (2025)
8. Taiti, C., Vivaldo, G., Mancuso, S., Comparini, D., Pandolfi, C.: Volatile organic compounds (vocs) fingerprinting combined with complex network analysis as a forecasting tool for tracing the origin and genetic lineage of arabica specialty coffees. Scientific Reports **15**(1) (2025)
9. Yafei, L., Chenlu, L., Xingang, L.: Long-term variations of ambient volatile organic compounds (vocs) from 2016 to 2020 in beijing, china (Jun 2023)
10. Yang, Y., Sun, F., Hu, C., Gao, J., Wang, W., Chen, Q., Ye, J.: Emissions of biogenic volatile organic compounds from plants: Impacts of air pollutants and environmental variables. Current Pollution Reports **11**(1) (2025)

# Performance Comparison of Torque Characteristics in Self-Excited Induction Generators for Three-Phase and Single-Phase Operation

Bruno Eduardo dos S. Romeiro[1,3], Francisco Ferreira Filho[2], Carlos Matheus R. de Oliveira[3], Cicero Hildenberg L. de Oliveira[3], and Ângela P. Ferreira[2]

[1] Instituto Politécnico de Bragança, Bragança, Portugal
[2] Research Centre in Digitalization and Intelligent Robotics (CeDRI), Laboratório Associado para a Sustentabilidade e Tecnologia em Regiões de Montanha (SusTEC), Instituto Politécnico de Bragança (IPB), Bragança, Portugal
apf@ipb.pt
[3] UTFPR, Universidade Tecnológica Federal do Paraná, IPCA, Brasil
carlosoliveira@utfpr.edu.br

**Abstract.** This study compares the performance of a self-excited three-phase induction generator (SEIG) in three-phase and single-phase (Fukami configuration) modes. The results show that three-phase operation delivers balanced voltages and stable torque, ensuring high efficiency and mechanical reliability. On the other hand, single-phase mode presents voltage imbalances and severe torque oscillations, which can damage the drive system due to mechanical stresses, reducing system efficiency and generation autonomy compared to three-phase operation. Capacitor sizing and dynamic simulations validated the analysis.

**Keywords:** Self-excited induction generator, Single-phase power, Torque stability.

## 1 Introduction

The recent blackout in the Iberian Peninsula, on 28 April 2025, left millions of people without an electricity supply and evidenced the vulnerability of current power infrastructures. The disruption affected telecommunications, transport systems and several other critical services. The time lapse till full restoration of the Spanish grid was about 23 hours, and the Portuguese grid was fully restored after 12 hours. In these circumstances, ensuring a rapid and effective emergency response remains essential, particularly in remote or vulnerable areas, and can be reliably achieved through the deployment of combustion engine-based generators [1].

Combustion engine-based generator systems can be designed using either synchronous or asynchronous generators, with the choice depending on several factors such as power requirements, efficiency, cost, and system complexity. Synchronous generators typically provide higher efficiency and superior power factor control, making them suitable for applications where precise voltage and frequency regulation are critical. However, they tend to be more expensive and mechanically complex due to the use of brushes, slip rings, and field windings. On the other hand, asynchronous generators, also known as induction generators, particularly those based on squirrel cage rotor designs, present several advantages that make them especially suitable for off-grid and emergency power applications. Their inherently simpler construction results in lower production and maintenance costs, while also enhancing mechanical

robustness and reliability. The feasibility of self-excitation of these machines for the voltage build-up process simplifies deployment and operation in remote or resource-constrained environments over synchronous generators [9].

In off-grid applications, the grid can be replaced by a properly dimensioned capacitor bank to enable self-excitation and autonomous operation [3]. Given that many essential loads, such as those in healthcare, communication, and water supply systems, require single-phase power, the adaptation of standard three-phase induction generators (IGs) for single-phase operation is a subject of considerable interest. Various strategies have been proposed and studied to enable reliable and efficient single-phase generation using conventional three-phase machines [4–6, 10].

While these strategies are well established in the literature, comprehensive assessments of their performance, particularly in comparison to conventional three-phase configurations, remain scarce. This study focuses on the behaviour of the electromagnetic torque. It is well known that torque ripple, especially when generators are driven by diesel engines, can produce various mechanical and operational challenges. These oscillations are not just a stability issue. Over time, they may accelerate mechanical wear and significantly reduce the service lives of generators and key components [2]. By comparing the three-phase and single-phase self-excited generator configurations under these conditions, this study aims to clarify the underlying torque dynamics in each scenario and assess their implications for practical applications in distributed generation.

## 2 Self-excited induction generator operation analysis

The magnetic cores of an induction generator possess residual magnetism, which is a prerequisite for generating electrical energy. In conjunction with the motion supplied by a prime mover, this magnetic field, reinforced by a reactive magnetising current supplied, is able to intensify the residual magnetism. For stand-alone operation (i.e., operation isolated from the primary grid), capacitor banks are employed to provide the magnetising current.

When a capacitor bank is connected in parallel with the machine's stator terminals, self-excitation takes place. The current in the excitation circuit flows through the stator winding and generates a magnetic flux in the same direction as the residual magnetic flux in the machine. This augments the residual flux and reinforces the magnetisation of the machine. The amount of capacitance determines how much voltage is generated [4].

In the proposed simulation, both the generator windings and the capacitor bank are connected in a star (wye) configuration, as depicted in Fig. 1. By maintaining an identical capacitance value $C_p$ in each phase, it is possible to generate balanced voltages $E_G$ within the system, thereby enabling the supply of balanced three-phase loads connected to the generator.

### 2.1 Induction machine especifications

The induction machine under study is a 4 hp three-phase squirrel-cage motor manufactured by WEG, model W22 Premium IR3 class, designed for high-efficiency

applications. It operates at 60 Hz with 4 poles, delivering a rated speed of 1745 rpm. The machine supports dual voltage operation (220/380 V) in delta/star configuration, with rated currents of 11.4 A and 6.61 A, respectively. The motor exhibits a power factor of 0.77, a service factor of 1.25, and a nominal efficiency of 89.5% at full load.

### 2.2 Three-phase operation

In Fig. 1 (a), a three-phase induction generator operates in isolated mode, self-excited by a capacitor bank with capacitance $C_p$ per phase. The dynamics of this configuration are inherently nonlinear due to the electromagnetic phenomena involved in the processes of self-excitation and voltage generation [3]. However, by employing suitable assumptions, it is possible to estimate with reasonable accuracy the required capacitance value in the capacitor bank to ensure that the machine enters the generating regime, given a predefined terminal voltage $V$ and operating frequency $f_b$ [8].

Let $Q_{C_p}$ denote the capacitive reactive power supplied by each phase of the capacitor bank connected in a wye configuration, and $Q_{\text{MAG}}$ the inductive reactive power required for the magnetization of the generator. Under steady-state conditions, the criterion for self-excitation is given by:

$$3Q_{C_p} = Q_{\text{MAG}} \tag{1}$$

At 60 Hz, the losses associated with modern capacitors can be neglected. Thus, the voltage across each capacitor can be estimated as:

$$V_{C_p} = X_{C_p} I_{C_p} \tag{2}$$

With this simplification, it becomes possible to determine the required capacitance $C_p$ such that the condition expressed in Equation 1 is satisfied:

$$\begin{aligned} Q_{cp} = V_{C_p}\ I_{C_p} &= \frac{Q_{\text{MAG}}}{3} = V_{C_p}\ \frac{V_{C_p}}{X_{C_p}} \\ X_{C_p} &= \frac{1}{2\pi f_b C_p} \end{aligned} \tag{3}$$

Leaving the capacitor voltage as a function of the line voltage $V_{C_p} = V_L/\sqrt{3}$, and rearranging Equation 3, we get,

$$C_p = \frac{Q_{\text{MAG}}}{2\pi f_b V_L^2} \tag{4}$$

The capacitance value obtained from Equation 4 represents only the minimum required capacitance, sufficient to supply the reactive power demanded by the magnetizing reactance of the machine. However, as the active power output increases, an additional amount of reactive power is required to sustain self-excitation and maintain a stable terminal voltage. To properly estimate this additional demand, the generator's behavior is simulated using the nameplate data and the equivalent circuit parameters, allowing for the determination of the optimal capacitor bank size needed to support nominal power generation [7]. Based on the specifications of the induction

machine under analysis, previously introduced, the minimum capacitance can then be calculated as follows [11]:

$$P_{\text{nom}} = 3 \text{ kW}, \quad \cos(\phi) = 0.77 \Rightarrow \phi = 39.65^\circ, \quad S_{\text{in}} = \frac{P_{\text{nom}}}{\cos(\phi)} = 3.8961 \text{ kVA},$$

$$Q_{\text{MAG}} = S_{\text{in}} \tan(\phi) = 3.229 \text{ kvar}, \quad Q_{C_p} = \frac{Q_{\text{MAG}}}{3} = 1.076 \text{ kvar},$$

$$C_{p\,\text{mín}} = \frac{Q_{C_p}}{V_{Cp}^2 \cdot 2\pi f} = \frac{1.076 \times 10^3}{220^2 \cdot 2\pi \cdot 60} = 58.9784 \ \mu\text{F} \tag{5}$$

Approximating the calculated value to the nearest commercially available capacitor, a value of $C_p = 60$ $\mu$F was adopted.

### 2.3 Single-phase operation - Fukami connection

In this work, we investigate the efficiency of employing a single-phase self-excited induction generator (SEIG) based on a modified induction motor (MIT), as configured in Fig. 1(b) [4]. This approach eliminates the need for complex voltage control systems, resulting in a highly cost-effective and reliable solution.

As detailed in [4], the values of the capacitors $C_s$ and $C_p$ are selected to minimize voltage regulation, with $C_s$ set to be 62.6% greater than $C_p$. Based on the capacitance value computed in Section 2.2, $C_p = 58.9784$ $\mu$F, we obtain $C_s = 95.8999$ $\mu$F. The closest commercially available capacitor value is $C_s = 100$ $\mu$F.

The generator setup comprises a three-phase squirrel-cage induction machine connected in a wye (Y) configuration. The capacitors $C_s$ and $C_p$ are arranged in a series-parallel network with a single-phase load. As shown in Fig. 1(b), all three stator windings are used both for excitation and power delivery, ensuring effective utilization of the magnetic core and maintaining system symmetry.

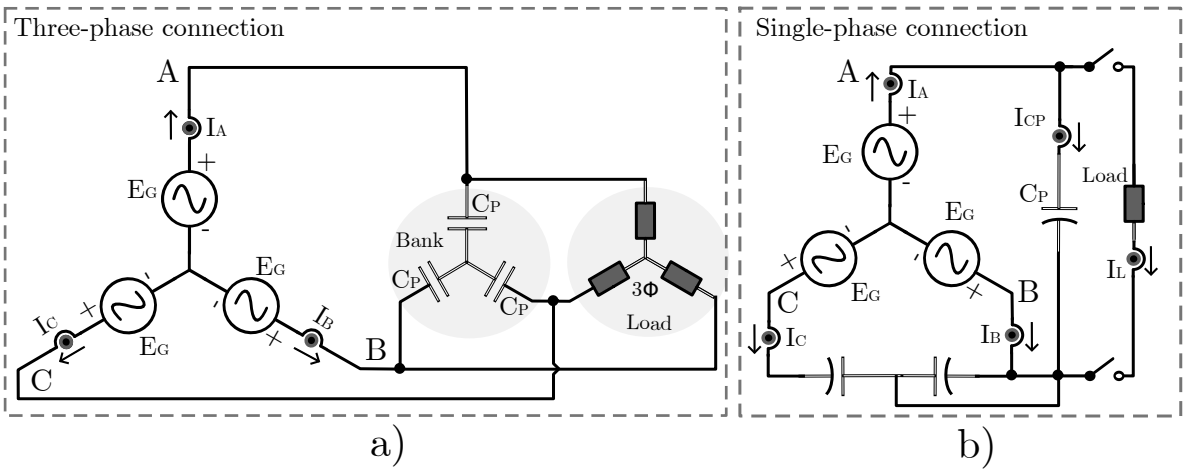
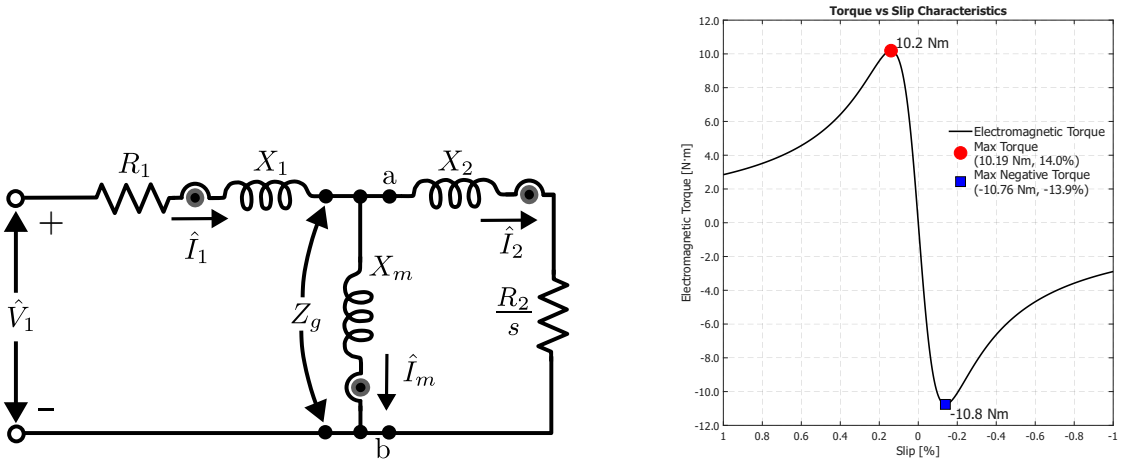



**Fig. 1.** Performance characteristics of the induction machine: a) Three-Phase Induction Generator, b) Three-Phase Induction Generator with Single-phase Operation

## 3 Torque and power

The analysis of the torque versus slip characteristic is fundamental in studying induction generators. This relationship makes it possible to identify the operating point at which the machine achieves optimal performance as a generator, namely, the point that provides the highest efficiency in converting mechanical energy into electrical energy. Understanding this curve enables adjustment of the load conditions and rotor speed to optimise system operation, minimise losses, and ensure greater stability in power generation, particularly in applications where the generator operates in isolation from the primary grid. The equivalent circuit of the induction machine, neglecting core losses, is shown in Fig. 2a.




**(a)** Equivalent Circuit

**(b)** Torque vs Slip Characteristics

**Fig. 2.** Performance characteristics of the induction machine

The Thévenin theorem can be applied to the machine's equivalent circuit to analyse torque and power. Considering a simple voltage divider, the following equations are obtained:

$$V_{1,\text{eq}} = V_1 \left( \frac{jX_M}{R_1 + j(X_1 + X_M)} \right), \tag{6}$$

$$Z_{1,\text{eq}} = R_{1,\text{eq}} + jX_{1,\text{eq}} = \frac{jX_M(R_1 + jX_1)}{R_1 + j(X_1 + X_M)}, \tag{7}$$

$$I_2 = \frac{V_{1,\text{eq}}}{\sqrt{\left(R_{1,\text{eq}} + \frac{R_2}{s}\right)^2 + (X_{1,\text{eq}} + X_2)^2}}, \tag{8}$$

$$\tau_{\text{mec}} = \frac{P_{\text{mec}}}{\omega_r} = \frac{P_{\text{EF}}}{\omega_{\text{sm}}}, \tag{9}$$

$$\tau_{\text{mec}} = \frac{1}{\omega_{\text{sm}}} \left[ \frac{N_{\text{fases}} V_{1,\text{eq}}^2 \cdot \frac{R_2}{s}}{\left(R_{1,\text{eq}} + \frac{R_2}{s}\right)^2 + (X_{1,\text{eq}} + X_2)^2} \right] \tag{10}$$

Fig. 2b defines the electromechanical torque versus slip curve of the induction machine used in this study. This curve characterises the dynamic behaviour of input power and identifies the typical operating point for isolated systems. The adopted operation point is below 20% of the maximum power output. Consequently, the common operating region for this generator is evaluated within a slip range of 10% to 14%, with torque values under 10 N·m.

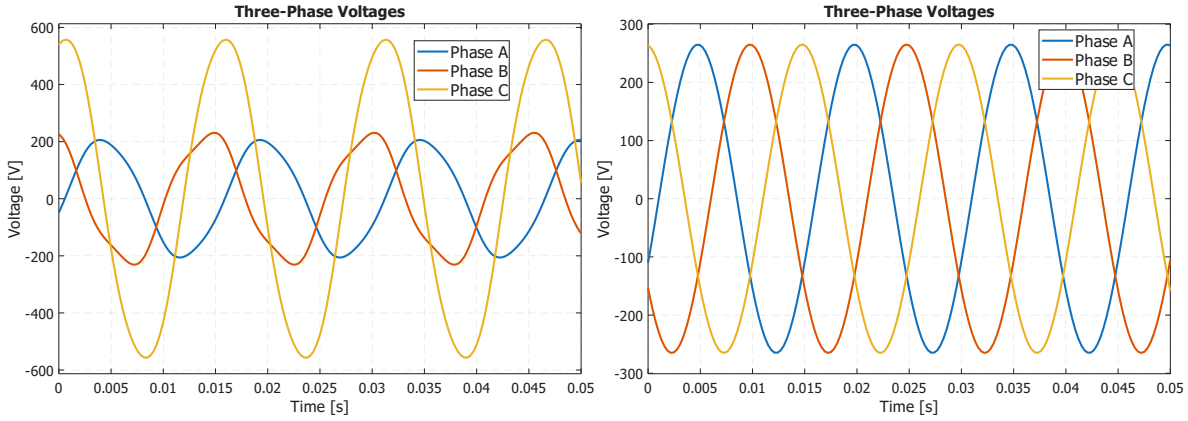
## 4 Results

It was decided to consider the same mechanical input power for both cases studied, ensuring comparable analysis conditions. It was observed that, when the operating regime approaches the maximum torque point of -10.76 N·m, as illustrated in Fig. 2b, the TIM enters a demagnetization regime, caused by the variation of the magnetization inductance considered in the simulation. To overcome this contingency, a torque of 6.8 N·m was defined as a reference, presenting a safe margin in relation to the critical region of the torque curve.

The rotation speed was kept fixed at 2050 rpm, seeking a negative slip of approximately -13.9%, calculated considering the characteristics of the TIM under study. Through the defined parameters – torque of 6.8 N·m and rotation of 2050 rpm, the mechanical input power $P_{\text{mec}} = \tau_{\text{mec}} \cdot \omega_r = 1459.8$ W was obtained.

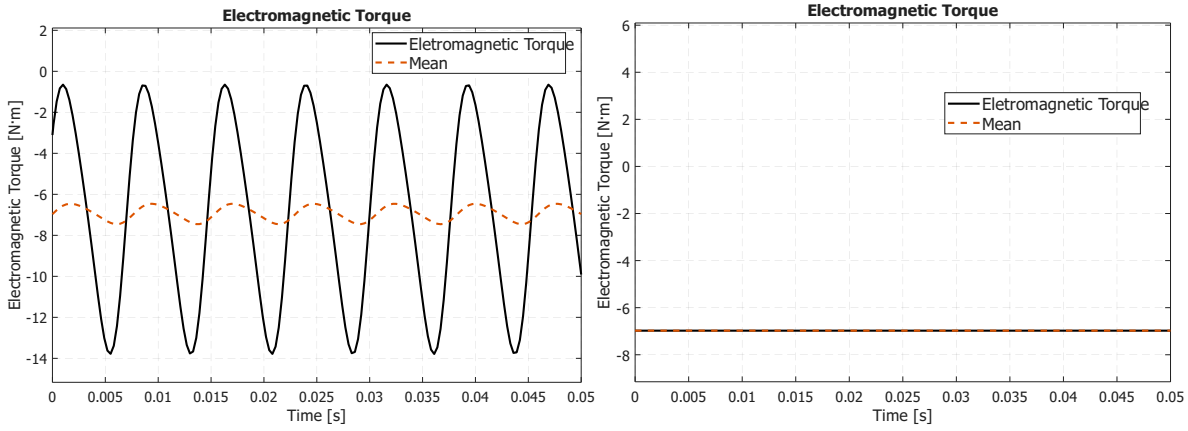
Based on the mathematical model presented in the previous section, the simulation results obtained using MATLAB/Simulink are presented below, allowing the analysis of the induction generator's performance under two different configurations: Case 1, Single-phase connection and Case 2, Three-phase connection.

Fig. 3 shows the behavior of the three-phase voltages generated in the two cases under analysis. For Case 1 (Fig. 3a), the phase waveforms present asymmetries and distortions, evidencing voltage unbalance and harmonic interference. In contrast, in Case 2 (Fig. 3b), the voltages of the three phases are clearly balanced, with pure sinusoidal shapes, evidencing an ideal condition of three-phase generation. In Fig. 4, the analysis of the electromagnetic torque on the machine shaft clearly shows the mechanical stresses involved in Case 1. In Fig. 4a, the torque presents a significant oscillation over time, with variations around the average value of 6.8 N·m. This ripple can cause mechanical vibrations and drastically reduce the efficiency of the system. In contrast, in Fig. 4b, corresponding to three-phase operation, the torque is practically constant and linear, indicating a more efficient and stable energy conversion.




**(a)** Single-phase Connection. **(b)** Three-phase Connection.

**Fig. 3.** Terminal Voltage Generator Comparison




**(a)** Single-phase Connection. **(b)** Three-phase Connection.

**Fig. 4.** Eletromagnetic Torque Comparison.

## 5 Conclusion

The results obtained clearly demonstrate the technical and operational limitations of TIMs operating in single-phase mode. The presence of strong torque oscillations and the evident imbalance in the generated voltage compromise not only the efficiency of the system, but also the mechanical durability of the equipment [2]. These undesirable conditions can lead to excessive vibrations, heating and reduced system autonomy.

On the other hand, operation in three-phase mode demonstrated significantly superior performance, with balanced voltage generation and constant torque. This torque stability is a clear indication of the advantage of three-phase operation, reducing mechanical stress and increasing machine reliability.

## References


1. O apagão ibérico de 2025: Vulnerabilidades sistémicas e lições para a resiliência energética europeia resumo. Tech. rep., A Pátria - Jornal da Comunidade Científica de Língua Portuguesa (4 2025)
2. van Binsbergen, D., Valavi, M., Nejad, A.R., Helsen, J.: Effect of generator torque ripple optimization on a geared wind turbine drivetrain. Forschung im Ingenieurwesen/Engineering Research **87**, 197–205 (2023). https://doi.org/10.1007/s10010-023-00624-3
3. Bodson, M., Kiselychnyk, O.: Nonlinear dynamic model and stability analysis of self-excited induction generators. In: Proceedings of the 2011 American Control Conference. pp. 4574–4579 (2011). https://doi.org/10.1109/ACC.2011.5991253
4. Fukami, T., Imamura, M., Kaburaki, Y., Miyamoto, T.: A new self-regulated self-excited single-phase induction generator using a squirrel cage three-phase induction machine **1**, 308–312 vol.1 (1995). https://doi.org/10.1109/EMPD.1995.500744
5. Fukami, T., Kaburaki, Y., Kawahara, S., Miyamoto, T.: Performance analysis of a self-regulated self-excited single-phase induction generator using a three-phase machine. IEEE Transactions on Energy Conversion **14**(3), 622–627 (1999). https://doi.org/10.1109/60.790925
6. Mahato, S.N., Sharma, M.P., Singh, S.P.: Determination of minimum and maximum capacitances of a self-regulated self-excited single-phase induction generator using a three-phase winding. In: India International Conference on Power Electronics. pp. 28–33 (2006)
7. Silva, E.O., Vanço, W.E., Guimarães, G.C.: Capacitor bank sizing for squirrel cage induction generators operating in distributed systems. IEEE Access **8**, 27507–27515 (2020). https://doi.org/10.1109/ACCESS.2020.2971704
8. Simone, G.A.: Máquinas de Indução Trifásicas – Teoria e Exercícios. Érica, Rio de Janeiro, 2 edn. (2009)
9. Singh, M., Singh, S.P., Singh, B., Pandey, A.S., Dixit, R., Mittal, N.: Stand alone power generation by $33\phi$ asynchronous generator: A comprehensive survey. In: 2012 2nd International Conference on Power, Control and Embedded Systems. pp. 1–14 (2012). https://doi.org/10.1109/ICPCES.2012.6508085
10. Smith, O.J.M.: Three-phase induction generator for single-phase line. IEEE Transactions on Energy Conversion **EC-2**(3), 382–387 (1987)
11. Syukri, M., Syuhada, A., Ramadhani, S., Affan, M., Yanis, M., et al.: Analysis of the effect of capacitors on the voltage generated by a 3-phase induction generator. In: 2022 International Conference on Electrical Engineering and Informatics (ICELTICs). pp. 173–177. IEEE (2022)

## Sponsors


– Fundação para a Ciência e a Tecnologia, FCT


– Projeto STEP (project n.: 101078933, has been funded by the European Commission)