ICAI 2025 Proceedings

Covilhã, 9-11 July, 2025

Contents

Keynote

Keynote speech by Professor Francesco Berganadano, from the Università di Torino.

Evasion resistance via diversity prediction

Francesco Bergadano
Dipartimento di Informatica
Università di Torino, Italy
francesco.bergadano@unito.it

Sandeep Gupta
Centre for Secure
Information Technologies
Queen's University Belfast, UK
s.gupta@qub.ac.uk

Bruno Crispo
Dipartimento di Ingegneria
e Scienza dell'Informazione
Università di Trento, Italy
bruno.crispo@unitn.it

***Index Terms*—Adversarial evasion, Randomization, Keyed learning, Moving Target Defense, Diversity prediction**

I. Extended abstract

Security events, with potentially rapid and destructive impact on ICT infrastructures, software and data, can be analyzed and categorized by automated AI systems, as an important part of defensive technologies. In many cases such AI defenses include classifiers, learned from previous data, that can label a current event as either acceptable or potentially malicious. This situation can be simplified as in in Fig. 1. When an attack is detected, there is a penalty for the adversary: an incident management system is triggered, and consequences can follow, going from IP address blocking, to honeypot activation, or even legal action when possible and appropriate.

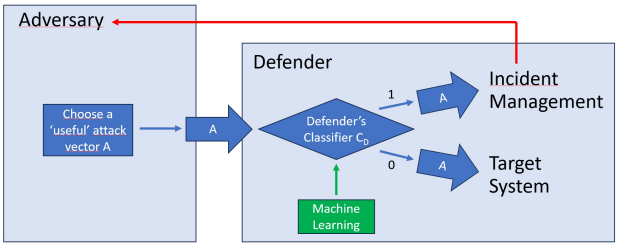


Fig. 1: Classifier-based security protection

The effectiveness of such defensive systems can be challenged by adversarial action (see, e.g., [1], [10], [13], [14], [21], [23], [24]), either during the training phase (*poisoning attacks*), or the inference phase (*evasion attacks*), causing incorrect classifications and making the target system reachable and potentially vulnerable. We address pure evasion attacks [2], [4], [8], [15], [21], [22], where the adversary tries to avoid detection by choosing stealthy attack patterns, but does not try to compromise the defender's learning phase.

An evasive adversary can mimic the defender's learning phase, and obtain a classifier that is identical or similar. The simple situation of Fig. 1 will then evolve into the one reported in Fig. 2. This attack methodology can be particularly dangerous for an AI defense, because it makes the filter ineffective, as the adversary can choose attack patterns that will be undetected and reach the target system.

A common strategy used to counter such an evasive adversary is *randomization* [3], [8], [9], [16], [17], [19], [22], where

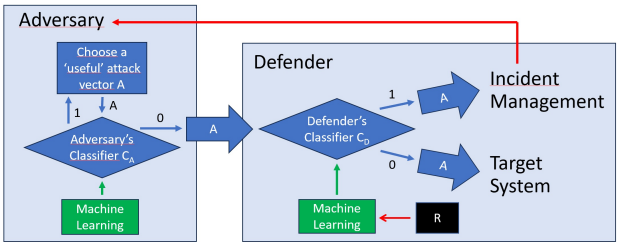


Fig. 2: Defensive system with an evasive adversary

random input is provided to the defender's Machine Learning component, making the classifier C_D difficult to predict or simulate for the adversary (input "R" in Fig. 2). This approach may have limitations, as it can prove ineffective [5], [11] or degrade accuracy [9], [22]. Goodfellow [11] states that:

> *An intriguing aspect of adversarial examples is that an example generated for one model is often misclassified by other models, even when they have different architecures or were trained on disjoint training sets. Moreover, when these different models misclassify an adversarial example, they often agree with each other on its class.*

We believe that randomization is indeed a powerful tool against adversarial evasion, but the threat model where it operates and the precise methodologies to be adopted need to be clarified [6]. First of all, we need to understand what kind of information is available to the adversary, and we would like to propose a "keyed box" approach, as described in Fig. 3.

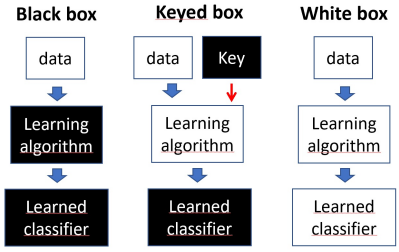


Fig. 3: Threat models: white rectangles are known to the adversary, while black rectangles are opaque

Using this keyed model of adversarial knowledge, we observe that the adversary cannot be sure of producing the same

classifier as the defender, because different inputs will be used, as detailed in Fig. 4. As a consequence, the adversarial evasion strategy of Fig. 2 will not necessarily achieve its goals. In particular, the adversary will fail, in two cases [7]:

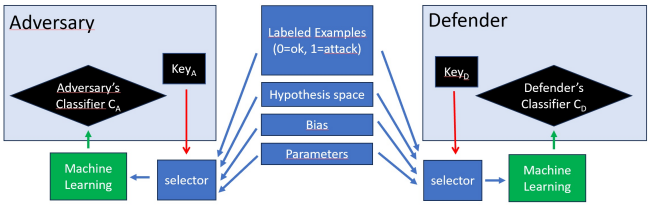


Fig. 4: Evasive adversary with keyed learning

1) no attack pattern A is found where $C_A(A) = 0$, i.e. the adversary cannot find an attack that, based on his knowledge, is likely to evade detection. In this case the adversary must refrain from attacking, as there are high chances of being caught and activating the incident management process.
2) $C_A(A) = 0$ but $C_D(A) = 1$, and the attacker will attempt attack A, erroneously thinking it will evade defender detection. Incident management will activate and consequences will follow.

Case 1 is unlikely in practical situations, where the target systems are very complex, and a perfect attack detector is unlikely to be found. We would then like to address case 2, where the adversary fails because C_A and C_D are dissimilar, in the following sense[1]:

Definition 1. ***Classifier diversity***
Consider two classifiers $C1$ and $C2$ and a set of examples E. Then their diversity D(C1,C2,E) is defined as follows:

$$D(C1, C2, E) = 1 - \frac{errors(C1, E) \cap errors(C2, E)}{errors(C1, E) \cup errors(C2, E)} \quad (1)$$

Classifier diversity can be extended to a set of classifiers:

Definition 2. ***Classifier diversity for a hypothesis space C***
Consider a set of possible classifiers C, and a set of examples E. Then their diversity D(C,E) is defined as follows:

$$D(C, E) = 1 - \frac{\cap_{C_i \in C} errors(C_i, E)}{\cup_{C_i \in C} errors(C_i, E)} \quad (2)$$

We must, however, specify what actually corresponds to the set of examples E of Definition 1, in the context of the evasion attack of Figures 2 and 4. We would like the *current* attack vector A to belong to E, because in that case we can infer that if the classifiers C_A and C_D have high diversity, it is possible that they will classify A differently and make the adversary fail. However, the attack A is new and the defender has no control over it. The best the defender can do is observe classifier diversity with respect to the training set Tr, and inductively assume that a similar diversity will be observed on future events, including the current attack vector A. In other words, the defender can *predict* classifier diversity, and use a hypothesis space where this prediction is as high as possible. We would then like to address the following prediction task:

Definition 3. ***Diversity prediction***
Consider a set of possible classifiers C, and a set of training examples Tr. We can then compute the diversity $D(C, Tr)$. Consider then an independent test sect Ts. Diversity can be predicted if $D(C, Ts)$ is, in the average case, sufficiently close to $D(C, Tr)$.

This can be approached by means of Hoeffding-style bounds, and Vapnik's theory of uniform convergence [20]:

$$Prob(|D(C, Tr) - D(C, Ts)| > \epsilon) < f(C, |Tr|, \epsilon) \quad (3)$$

Such bounds are often not tight when it comes to the usual case of accuracy prediction [18], and hence of little practical use. The case of diversity prediction can only be more difficult because it involves many classifiers and their error correlation, and not only their performance.

As future work we would then like to address diversity prediction experimentally: split available data sets into training and test subsets, and measure $|D(C, Tr) - D(C, Ts)|$, in order to understand if this prediction task can be targeted in practice. If the answer is positive, many downstream cybersecurity applications may be approached in order to resist evasion, following the *keyed* learning defense described in this paper.

[1]This is an adaptation of a measure known as Jaccard similarity [12]

Acknowledgments

This project was partially funded by the project SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union – NextGenerationEU, specifically by the cascate call project Q-CPS2: Quantitative models for Cyber Physical Systems Security. This manuscript reflects only the authors' views and opinions and the Ministry cannot be considered responsible for them. Partial funding was also provided by the European Union under NextGenerationEU, PRIN 2022 Prot. n. 202297YF75.

References

[1] Alsmadi, I.: Adversarial machine learning, research trends and applications. In: Baddi, Y., Gahi, Y., Maleh, Y., Alazab, M., Tawalbeh, L. (eds.) Big Data Intelligence for Smart Applications, pp. 27–55. Springer (2022). https://doi.org/10.1007/978-3-030-87954-9_2
[2] Amich, A., Eshete, B.: Explanation-guided diagnosis of machine learning evasion attacks. In: Security and Privacy in Communication Networks. pp. 207–228. Springer (2021)
[3] Amich, A., Eshete, B.: Morphence: Moving target defense against adversarial examples. In: Proc. of the 37th Annual Computer Security Applications Conference. p. 61–75. ACM, New York, NY, USA (2021). https://doi.org/10.1145/3485832.3485899
[4] Apruzzese, G., Andreolini, M., Marchetti, M., Venturi, A., Colajanni, M.: Reinforcement adversarial learning against botnet evasion attacks. IEEE Trans. Netw. Serv. Management **17**, 1975–1987 (2020)
[5] Bakos, S., Madani, P., Davoudi, H.: Noise as a double-edged sword: Reinforcement learning exploits randomized defenses in neural networks (2024), https://arxiv.org/abs/2410.23870
[6] Bergadano, F.: Keyed learning: An adversarial learning framework. ETRI Journal **41**(5) (2019)

[7] Bergadano, F., Gupta, S., Crispo, B.: Keyed randomization with adversarial failure curves and moving target defense. In: Proceedings of the Fifth Intelligent Cybersecurity Conference (ICSC) (2025)

[8] Biggio, B., Corona, I., Maiorca, D., Nelson, B., Srndic, N., Laskov, P., Giacinto, G., Roli, F.: Evasion attacks against machine learning at test time. https://arxiv.org/abs/1708.06131v1 (2017)

[9] Biggio, B., Fumera, G., Roli, F.: Adversarial pattern classification using multiple classifiers and randomization. In: Int. W. on Structural, Syntactic and Statistical Pattern Recognition (2008)

[10] Dalvi, N., Domingos, P., Mausam, Sanghai, S., Verma, D.: Adversarial classification. In: Proc. ACM Int. Conf. on Knowledge Discovery and Data Mining (2004). https://doi.org/10.1145/1014052.1014066

[11] Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: 3rd Int. Conf. on Learning Representations, (ICLR), San Diego, CA (2015), http://arxiv.org/abs/1412.6572

[12] Jaccard, P.: Étude comparative de la distribution florale dans une portion des alpes et des jura. Bulletin de la Société Vaudoise des Sciences Naturelles **37**, 547–579 (1901)

[13] Lucas, K., Pai, S., Lin, W., Bauer, L., Reiter, M.K., Sharif, M.: Adversarial training for Raw-Binary malware classifiers. In: 32nd USENIX Security Symposium. pp. 1163–1180. Anaheim, CA (2023), https://www.usenix.org/conference/usenixsecurity23/presentation/lucas

[14] Mbow, M., Roman, R., Takahashi, T., Sakurai, K.: Evading iot intrusion detection systems with gan. In: 2024 19th Asia Joint Conf. on Information Security (AsiaJCIS). pp. 48–55 (2024). https://doi.org/10.1109/AsiaJCIS64263.2024.00018

[15] MITRE: Mitre atlas: Evading ml models. https://atlas.mitre.org/techniques/AML.T0015 (2024), accessed: 2024-12-24

[16] Mrdovic, R.S., Drazenovic, B.: Kids: a keyed intrusion detection system. In: Proc. DIMVA (2010)

[17] Nowroozi, E., Mohammadi, M., Golmohammadi, P., Mekdad, Y., Conti, M., Uluagac, S.: Resisting deep learning models against adversarial attack transferability via feature randomization. IEEE Trans. on Services Computing **17**(01), 18–29 (2024). https://doi.org/10.1109/TSC.2023.3329081

[18] Saitta, L., Bergadano, F.: Pattern recognition and valiant's learning framework. IEEE Transactions on Pattern Analysis and Machine Intelligence **15**(2), 145–155 (1993). https://doi.org/10.1109/34.192486

[19] Taran, O., Rezaeifar, S., Holotyak, T., Voloshynovskiy, S.: Machine learning through cryptographic glasses: combating adversarial attacks by key-based diversified aggregation. EURASIP J. on Inf. Security (2020)

[20] Vapnik, V.: Estimation of Dependences Based on Empirical Data. Springer (1982)

[21] Wang, S., Ko, R.K.L., Bai, G., Dong, N., Choi, T., Zhang, Y.: Evasion attack and defense on machine learning models in cyber-physical systems: A survey. IEEE Communications Surveys Tutorials **26**(2), 930–966 (2024). https://doi.org/10.1109/COMST.2023.3344808

[22] Yang, F., Chen, Z., Gangopadhyay, A.: Using randomness to improve robustness of tree-based models against evasion attacks. IEEE Transactions on Knowledge and Data Engineering **34**(2), 969–982 (2022)

[23] Zhang, H., Li, X., Tang, J., Peng, C., Wang, Y., Zhang, N., Miao, Y., Liu, X., Choo, K.K.R.: Hiding in plain sight: Adversarial attack via style transfer on image borders. IEEE Trans. on Computers **73**(10) (2024)

[24] Zhou, S., Liu, C., Ye, D., Zhu, T., Zhou, W., Yu, P.S.: Adversarial attacks and defenses in deep learning: From a perspective of cybersecurity. ACM Computing Surveys **55**(8), 1–39 (2022)

Session 1

Blockchain-Based Framework for Academic Credential Management: A Case Study at Guarda Polytechnic Institute

Pedro Pinto
Guarda Polytechnic Institute
Guarda, Portugal
ppinto@ipg.pt

Fátima Gonçalves
Guarda Polytechnic Institute
Guarda, Portugal
fgoncalves@ipg.pt

***Abstract*—This paper presents a blockchain-based framework for issuing and verifying academic certificates, offering an alternative to conventional centralized systems. Traditional certification methods are prone to fraud, administrative inefficiency, and interoperability limitations. Blockchain provides a decentralized architecture that enables immutability, transparency, and real-time credential validation. The proposed model integrates tokenization, digital signatures, and distributed storage to ensure data integrity and facilitate credential portability. Smart contracts are used to automate verification processes and manage access permissions. A case study carried out at Guarda Polytechnic Institute illustrates the technical feasibility of this approach, detailing its integration with institutional academic management systems. The implementation allows for the secure registration of diplomas, direct verification by third parties, and user-controlled access management. The results indicate a reduction in administrative workload, increased verification efficiency, and improved resistance to document tampering. The proposed solution complies with data protection regulations and supports cross-border recognition of academic qualifications.**

Keywords—Blockchain, Academic Certificates, Higher Education Institutions, Smart Contracts, Digital Credentials.

I. Introduction

Blockchain technology has demonstrated transformative potential across multiple sectors, including the issuance and verification of academic certificates [1]. Its decentralized, immutable, and transparent architecture addresses critical limitations of traditional certification systems, which are susceptible to forgery, rely on inefficient administrative procedures, and incur significant operational costs [2], [3].

In a blockchain-based framework, certificates are tokenized and converted into unique digital entities recorded on a distributed ledger. Digital signatures ensure documents authenticity and integrity while consensus mechanism such as Proof of Work (PoW), Proof of Stake (PoS), and Practical Byzantine Fault Tolerance (PBFT) establish trust and validate transactions without relying on centralized authorities [4], [5], [6]. The integration of smart contracts further automates the verification process, removing intermediaries and enabling students to control access to their academic records [7].

The core structure of blockchain, comprising sequential blocks linked through cryptographic hash functions, ensures record integrity and prevents post-validation modifications [8]. This architecture offers significant benefits in academic context, including enhance resistance to fraud, instant verification by third parties, and process transparency for all stakeholders, such as students, employers, and educational institutions [9].

Tokenization of academic credentials embeds metadata including student identity, qualification details, and learning chronology into verifiable digital assets [10]. This supports the construction of longitudinal academic portfolios and ensures record persistence and accessibility regardless of the issuing institution's availability [11], [12].

Pilot implementations and case studies confirm the technical and operational feasibility of blockchain in academic certification, highlighting its potential to align institutional certification systems with contemporary requirements for security, mobility, and verification efficiency [13], [14].

The main objective of this study is to design a blockchain-based model for issuing and verifying academic certificates and to validate its technical feasibility through a case study conducted at the Guarda Polytechnic Institute (IPG).

This paper is organized as follows: Section II reviews the related literature; Section III outlines the integration of blockchain technology into academic certification systems; Section IV details the case study conducted at IPG; and Section V presents the conclusions and discusses future research directions.

II. Literature Review

Academic certificates and diplomas are essential instruments for validating the knowledge and competencies acquired throughout an individual's educational trajectory. Their authenticity, however, is often difficult to verify, particularly in contexts of international academic and professional mobility [15]. In Portugal, two primary types of academic documents are issued: the certificate of qualifications and the diploma [16].

The certificate of qualifications attests to the completion of a specific level of education, such as basic, secondary, or higher education (bachelor's, master's, or doctoral degrees). It typically includes the holder's identification, level of education attained, academic transcript, individual course grades, and the final grade point average [17]. The diploma certifies the completion of a specific program, e.g. bachelor's in computer science or a short-cycle technical degree (CTeSP) in cybersecurity and may be accompanied by a diploma supplement providing additional information on the qualification level, curriculum, and learning outcomes [18]. These documents must be formally requested from the issuing institution, either in person or through digital platforms, depending on institutional procedures.

Traditional digital certificates, without blockchain integration, present technical and operational limitations. They are often issued through manual and managed using centralized systems, resulting in inefficiencies, susceptibility

to errors, and elevated risks of fraud [3], [14]. Despite including basic security mechanisms, conventional digital certificates remain vulnerable to forgery or unauthorized alterations [19].

Economically, the issuance and long-term maintenance of these certificates require centralized infrastructure, which increases operational costs [3], [19]. From a functional perspective, these systems present portability limitations, with challenges related to recognition and compatibility across different systems or jurisdictions, hindering international academic and professional mobility [19]. The revocation process for compromised certificates is typically bureaucratic and time-consuming [14], and many systems lack transparency regarding the issuance and verification lifecycle. Additionally, traditional certificates are susceptible to loss or destruction, posing risks to data security. These constraints have motivated the development of blockchain-based alternatives that improve the security, efficiency, and traceability of academic records [3].

The verification of certificate authenticity remains a major challenge for employers and public entities. Serranito, Vaconcelos, Guerreiro and Correia [14] propose a blockchain-based solution using smart contracts to enable decentralized validation of higher education certificates. In their prototype, diplomas are registered on a blockchain, and their authenticity can be verified independently by third parties. Similarly, Mikroyannidis, Domingue, Bachler, and Quick [20] propose a peer-to-peer infrastructure for education, in which blockchain supports the management and accreditation of learning experiences, in formal and informal contexts, granting learners full control over their credentials [21].

A. Integration of Blockchain into Academic Certification Systems

Blockchain technology constitutes a significant advancement in digital systems architecture, providing a decentralized and secure framework for issuing and verifying academic credentials. Its structure as a distributed ledger, composed of interconnected data blocks linked sequentially through cryptographic hashes, ensures data integrity, immutability, and traceability. These features are particularly relevant in academic contexts, where resistance to tampering and reliable validation of records are essential [22].

Blockchain falls within the domain of Distributed Ledger Technology (DLT), establishing a system for recording, sharing, and synchronizing data across multiple devices or locations without the need for centralized entities to validate or manage the information [23]. This model represents a departure from traditional academic certificate management systems, which are based on centralized and hierarchical control by issuing institutions and are often subject to complex and time-consuming bureaucratic procedures, as mentioned earlier. By implementing blockchain-based solutions, higher education institutions (HEIs) can overcome the inherent limitations of current certification systems, particularly the risks of forgery and the challenges associated with international credential verification.

The fundamental structure of blockchain consists of blocks composed of three core components [24], [25]. The data component may represent detailed information about qualifications, acquired competencies, and the student's educational path in an academic context. The hash, a unique identifier of the block, functions as a cryptographic fingerprint that ensures the certificate's authenticity. The previous block hash establishes the sequential link between blocks, thereby preserving the chronological integrity of the documented academic trajectory. This architecture endows the system with properties of transparency, security, and resistance to forgery, as any attempt to alter a block would require modifying all subsequent blocks—a computationally infeasible operation due to the distributed nature of the network.

The process begins when a user submits a transaction to the network, which is then aggregated into a block along with other pending transactions (Figure 1). This candidate block is propagated across all network nodes, which proceeds to validate it using specific consensus algorithms. The PoW mechanism, for instance, requires validator nodes (known as miners) to solve complex cryptographic problems, consuming significant computational resources to demonstrate their commitment to maintaining the integrity of the network [22]. Alternatively, PoS protocols base validation on the amount of cryptocurrency held by the validator, significantly reducing the energy consumption associated with the process [26]. Once consensus is achieved, the block is added to the chain, becoming a permanent part of the distributed ledger, and the validator nodes are rewarded according to the specific blockchain protocol in use.

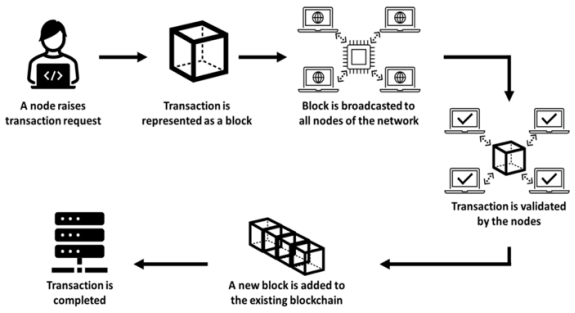


Fig. 1. Transaction flow in a blockchain network

Source: Santhi and Muthuswamy [33].

Smart contracts represent an evolution in the blockchain paradigm, transcending the basic transactional model by implementing self-executing programmable logic within the distributed infrastructure. Conceptually, a smart contract is a computational protocol that autonomously facilitates, verifies, and enforces the negotiation or execution of a contract without the need for intermediaries [26]. From a technical perspective, smart contracts are implemented using specific programming languages - such as Solidity for the Ethereum platform - and are compiled and deployed onto the blockchain through transactions. Once deployed, they become autonomous entities with their address, capable of interacting with other contracts and external entities through well-defined interfaces [27]. Smart contracts are applied across various domains, including decentralized finance (DeFi), logistics, insurance, digital identity, and decentralized governance [25]. In the context of academic certificates, they enable the implementation of automated issuance and verification systems, eliminating the need for intermediaries and significantly reducing the risk of fraud.

An analysis of the fundamental properties of blockchain technology reveals a set of distinctive attributes that enhance its applicability in the field of academic certification. Zeng, Xie, Dai, Chen and Wang [25] list important features, such as

decentralization, which means that data from academic certificates can be checked across multiple nodes in the network, without the need for institutions to act as middlemen in the verification process. Immutability, as once a certificate is recorded, cannot be altered or forged, directly addressing the issue of document fraud discussed earlier. Transparency is achieved by ensuring that all participating institutions in the network have access to the same certification history, thereby facilitating inter-institutional recognition of qualifications. Cryptographic security, which employs advanced cryptographic mechanisms to ensure the integrity and authenticity of diplomas, significantly outperforms conventional security methods.

The operation of blockchain technology in the context of academic certification relies on specific structural and procedural elements. According to the research conducted by Christidis and Devetsikiotis [28], key components include: blocks, which in the educational domain may contain complete diploma records, including course units, grades, and acquired competencies; transactions, which may represent the issuance of new certificates or the validation of existing credentials; the distributed ledger, which maintains the complete history of academic certifications and is accessible to authorized entities such as educational institutions, employers, and government agencies; and cryptographic algorithms, which ensure the authenticity and integrity of academic certificates, effectively eliminating the risks of forgery.

In the absence of a central evaluating authority, blockchain consensus mechanisms serve as essential protocols for synchronization and agreement among participating academic institutions. Mingxiao, Xiaofeng, Zhe, Xiangwei and Qijun [29] say that PoW, PoS, and PBFT are the main consensus models that can be used for academic certification. Each one has its own pros and cons when it comes to speed, security, and scalability when validating credentials. To validate transactions and achieve consensus in a decentralized network, PoW leverages computational power to solve complex mathematical problems. This mechanism is fundamental to Bitcoin's operation and was historically used by Ethereum. PoS, on the other hand, relies on validators who stake crypto assets as collateral to validate transactions and reach consensus in a decentralized environment. This mechanism is employed by many newer blockchains, including Ethereum, Polkadot, Cardano, and Solana. PBFT is a consensus algorithm that depends on a set of known validators to achieve consensus. Validators in a blockchain are nodes or participants responsible for verifying, validating, and recording transactions within the blockchain network, thereby helping maintain consensus and system integrity.

According to Wang, Han, and Beynon-Davies [6], some educational implementations of blockchain adopt hybrid consensus models that combine features of PoW and PoS. These models allow accredited academic institutions to participate as privileged validators of certificates, preserving institutional authority in the certification process while benefiting from the advantages of decentralization.

The versatility of blockchain technology has catalyzed its implementation across multiple educational sectors, complementing and transforming certification systems. As documented by Casino, Dasaklis, and Patsakis [30], current educational applications include the issuance of verifiable diplomas, addressing previously identified portability issues; the creation of lifelong learning passports, integrating certifications from multiple institutions into a single verifiable record; the automation of qualification verification by potential employers, eliminating the time-consuming processes already described; and the cross-border recognition of academic qualifications, directly addressing the limitations of current international recognition systems.

Grech and Camilleri [3] have also identified significant benefits of implementing blockchain in education, particularly in addressing the challenges of transnational academic certification. Among the highlighted benefits are the permanence of records regardless of the issuing institution's future, the elimination of the need for centralized verification bodies, the transferability of certificates across countries and educational systems, and the empowerment of students as lifelong owners and managers of their credentials.

The integration of blockchain technology with current academic certification systems presents significant challenges related to scalability, interoperability, and institutional acceptance. Turkanović, Hölbl, Košič, Heričko and Kamišalić [7] propose the EduCTX model, a blockchain-based platform adapted to the European Credit Transfer and Accumulation System (ECTS), demonstrating the feasibility of transforming traditional certification processes. This model offers a solution for the globalization of higher education by enabling the immediate transferability and verifiability of academic credits between institutions, overcoming the bureaucratic barriers identified in current certification systems.

Academic literature on blockchain applied to academic certification has been prolific in recent years, addressing not only technological aspects but also pedagogical, institutional, and regulatory implications. Williams [31] examines the potential of blockchain to democratize access to education through decentralized certification, offering alternatives to hierarchical systems. Jirgensons and Kapenieks [9] explore the use of blockchain to create a certification system resilient to forgery, directly addressing the vulnerabilities of traditional certificates. Ocheja, Flanagan, and Ogata [11] propose a framework for interoperability among different educational blockchain systems, facilitating academic mobility beyond the current limitations of cross-border recognition.

Governance and revocation mechanisms are also critical for ensuring long-term trust and institutional adoption of blockchain-based certification systems [34], [35], [36].

III. Case Study Description

The implementation of a blockchain-based solution for academic certification at IPG could represent a significant advancement in the authentication and management of diplomas within the institution. Following an approach similar to that of QualiChain, as described by Guerreiro, Ferreira, Fonseca, and Correia [13], this proposal suggests integrating a blockchain platform with IPG's academic management system, enabling secure and decentralized instant certificate validation.

Currently, HEIs face challenges in verifying the authenticity of diplomas - a process that can be time-consuming and bureaucratic. According to research by Guerreiro, Ferreira, Fonseca, and Correia [13], combining an academic system with blockchain technology would allow the recording of a cryptographic hash for each diploma. This would make sure that the diploma can't be changed and let employers and other outside groups check its legitimacy without having to contact the institution that issued it directly.

The solution to be implemented at IPG would follow a similar approach, whereby diplomas would be automatically digitally signed and recorded on the blockchain, ensuring their authenticity, integrity, and compliance with the General Data Protection Regulation (GDPR). In this way, students would retain control over access to their certifications, with the ability to share or revoke permissions as needed.

The proposed model would involve the academic registry office at IPG, responsible for recording certificates in the system; graduating students, who would be able to manage access to their diplomas; and employers, who could instantly validate the authenticity of the documents. The adoption of this model could reduce academic fraud, accelerate recruitment processes, and enhance the credibility of the certification issued by the IPG.

Furthermore, the experience reported in the study by Guerreiro, Ferreira, Fonseca, and Correia [13] suggests that users value this type of solution for its transparency, ease of access, and security. Based on these results, IPG would be able to test how blockchain technology affects the speed and trustworthiness of administrative tasks and the certification process for academics. This would create a new model that fits with global trends in the digitalization of higher education.

A. Stakeholders and use cases

The digital certification process at IPG would involve multiple stakeholders (Figure 2). The main actors in this ecosystem would be graduating students, responsible for managing access to their diplomas; the IPG academic registry office, in charge of issuing and recording credentials; and employers and public sector entities, who would be able to autonomously verify the authenticity of diplomas.

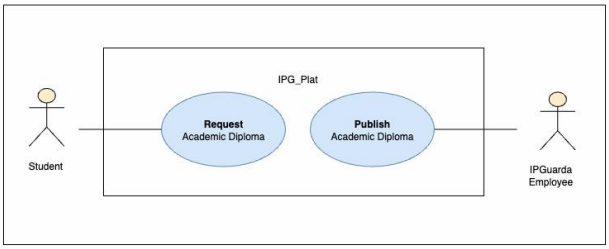


Fig. 2. Use cases of the IPG Blockchain system.

Source: Adapted from Guerreiro, Ferreira, Fonseca, and Correia [13].

The information flow of this system, represented in Figure 3, would begin with the diploma request submitted by the graduating student. Next, the IPG academic registry office would issue and digitally store the diploma, recording a cryptographic hash of the document on the blockchain. The graduate would have the ability to authorize and revoke access to their diploma, granting permission to employers and interested institutions. The authenticity of the diploma would be verified by consulting the hash recorded on the blockchain, eliminating the need for direct intervention from IPG in the credential verification process.

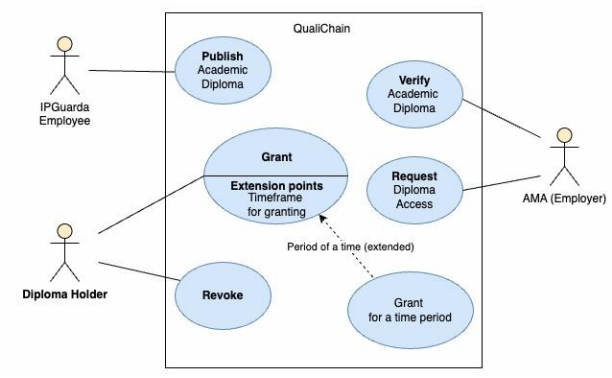


Fig. 3. Use cases of the QualiChain system.

Source: Adapted from Guerreiro, Ferreira, Fonseca, and Correia [13].

Figure 4 presents the information flow between the different actors in the system, using the Business Process Model and Notation (BPMN). In this model, the IPG academic registry office plays a central role in publishing and recording diplomas, while graduates retain control over access to their certificates. The authentication process occurs in a centralized manner, ensuring security, immutability, and transparency in the validation of academic credentials. The process begins when a student requests the issuance of their diploma upon completion of their academic program, whether in a bachelor's, master's, or postgraduate course. In response to this request, the registry office processes the application and digitally generates the diploma, storing it in the institution's academic management system. This diploma, initially made available in PDF format, is stored on IPG's servers for future consultation.

The innovation introduced by integrating blockchain into this process lies in the next step: IPG's academic system would automatically generate a unique hash for each issued diploma, ensuring its inviolability and authenticity. This cryptographic hash would be stored on a blockchain, guaranteeing that any modification or forgery attempt could be easily detected. Once recorded on the blockchain, the diploma holder gains full control over their digital document. Through a secure platform, graduates would be able to share their diploma with employers, public entities, or other educational institutions, providing them with direct access to credential validation.

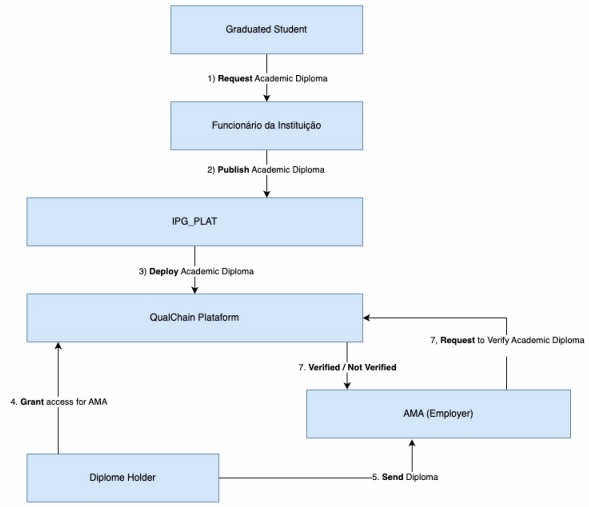


Fig. 4. Information flow for a verified and unverified diploma.

Source: Adapted from Guerreiro, Ferreira, Fonseca, and Correia [13].

The verification process would be carried out in a decentralized manner by querying the blockchain to confirm whether the presented diploma matches the original record issued by IPG. Additionally, the platform would provide a flexible access management system, allowing the graduate to grant or revoke access permissions to the diploma whenever necessary. For example, a student who has shared their diploma with a potential employer could later remove that permission if it is no longer relevant. This blockchain-based model would bring significant benefits to IPG and all stakeholders involved in the academic certification process. For the academic registry, it would reduce the administrative burden associated with manual diploma authentication. It would provide graduates with a safe and trustworthy way to verify their credentials. For employers and external entities, it would streamline the recruitment process by eliminating the need to manually confirm the authenticity of documents with the issuing institution.

This work is situated at a conceptual and exploratory stage, presenting a technological proposal and its potential for practical implementation. However, it does not yet include empirical validation with real users.

IV. Conclusion

This study has analyzed the transformative potential of blockchain technology in the issuance and validation of academic certificates, with particular focus on a possible implementation at IPG. The results show that this technology is a strong and new alternative to traditional academic certification systems. It gets around problems like fraud, inefficient administration, and problems with verifying credentials across borders. The blockchain architecture - characterized by its decentralized, immutable, and transparent nature - shows strong potential to support a certification system with enhanced security, efficiency, and reliability. By implementing a blockchain-based system, IPG could ensure the permanent and tamper-proof registration of issued diplomas and certificates, granting students control over and the ability to share their credentials while enabling employers and other institutions to autonomously verify the authenticity of these documents.

The proposed case study illustrates the technical and operational feasibility of this implementation, highlighting benefits for all stakeholders involved in the academic certification process. For the institution, the simplification of administrative procedures and reduction of operational expenses represent significant advantages. Blockchain technology gives students control over their academic history and makes career mobility easier. For employers and external entities, the ability to instantly and autonomously verify certificate authenticity is an undeniable asset. Despite the challenges associated with implementing this technology - particularly the need to adapt existing systems and ensure compliance with data protection regulations - the empirical evidence presented suggests that the benefits substantially outweigh the difficulties. The experiences of the QualiChain and Fénix projects demonstrate that similar solutions have already been successfully implemented, reinforcing the viability of the proposed model.

Basically, adding blockchain technology to IPG's certification system is a smart move that will bring academic management up to date, boost the credibility of the institution, and meet modern needs for mobility and the ability to check credentials. In the current context of higher education digitalization and labor market internationalization, this technological innovation may serve as a significant competitive advantage, aligning the institution with the most advanced trends in educational credential management.

As future work, we intend to implement a pilot test involving students and employers to evaluate the system's usability, perceived trustworthiness, and impact on administrative efficiency.

References

[1] L. Palma, M. Vigil, F. Pereira and J. Martina, "Blockchain and smart contracts for higher education registry in Brazil" *International Journal of Network Management*, vol.29, no.3, pp. 20-61, 2019. Available: https://doi.org/10.1002/nem.2061

[2] Y. Noshi, "Development of blockchain-based academic credential verification system," *Open Access Library Journal*, vol. 11, no. 9, 2024. Available: https://doi.org/10.4236/oalib.1112130

[3] A. Grech and A. Camilleri, *Blockchain in education*. EU 278778 EN, Publications Office of the European Union, 2017. Available: https://dx.doi.org/10.2760/60649

[4] G. Caldarelli & J. Ellul, "Trusted academic transcripts on the blockchain: A systematic literature review," *Applied Sciences*, vol. 11, no.4, pp. 18-42, 2021. Available: https://doi.org/10.3390/app11041842

[5] J. Rooksby and K. Dimitrov, "Trustless education? A blockchain system for university grades," in *New Values Transactions: Understanding and Designing for Distributed Autonomous Organisations*, Workshop at DIS2017, June, Edigurgh, 2017. Available: https://johnrooksby.org/

[6] Y. Wang, J. Han, and P. Beynon-Davies, "Understanding blockchain technology for future supply chains: A systematic literature review and research agenda," *Supply Chain Management: An International Journal,* vol. 24, no. 1, pp. 62-84, 2019. Available: https://doi.org/10.1108/SCM-03-2018-0148

[7] M. Túrkanović, M. Hölbl, K. Košič, M. Heričko, and A. Kamišalić, "EduCTX: A blockchain-based higher education credit platform," *IEEE Access*, vol. 6, pp. 5112-5127, 2018. Available: https://doi.org/10.1109/ACCESS.2018.2789929

[8] D. Jose, J. Holme, A. Chakravorty, and C. Rong, "Integrating big data and blockchain to manage energy smart grids – TOTEM framework," *Blockchain: Research and Applications*, vol.3, no.3, 2022. Available: https://doi.org/10.1016/j.bcra.2022.100081

[9] M. Jirgensons and J. Kapenieks, "Blockchain and the future of digital learning credential assessment and management," *Journal of Teacher Education for Sustainability*, vol.20, no. 1, pp. 145-156, 2018. Available: https://doi.org/10.2478/jtes-2018-0009

[10] M. Han, Z. Li, J. He, D. Wu, Y. Xie, and A. Baba, "A novel-blockchain-based education records verification solution," in *Proc. 19th Annu. SIG Conf. on Information Technology Education*, 2018. Available: https://doi.org/10.1145/3241815.3241870

[11] P. Ocheva, B. Flanagan, and H. Ogata, "Connecting descentralized learning records: A blockchain-based learning analytics platform," in *Proc. 8th Int. Conf. on Learning Analytics and Knowledge*, 2018, pp. 265-269. Available: https://doi.org/10.1145/3170358.3170365

[12] H. Sun, X. Wang, and X. Wang, "Application of blockchain technology in online education, "*International Journal of Emerging Technologies in Learning (iJET)*, vol. 13, no. 10, pp. 252-259, 2018. Available: https://doi.org/10.3991/ijet.v13i10.9455

[13] S. Guerreiro, J. Ferreira, T. Fonseca, and M. Correia, "Integrating an academic management system with blockchain: A case study," *Blockchain: Research and Applications*, vol. 3, no. 4, 2022. Available: https://doi.org/10.1016/j.bcra.2022.100099

[14] D. Serranito, A. Vasconcelos, S. Guerreiro, and M. Correia, "Blockchain ecosystem for verifiable qualifications," in *Proc. 2nd Conf. on Blockchain Research & Applications for Innovative Networks and Services (BRAINS)*, Paris, 2020, pp. 192-199. Available: https://doi.org/10.1109/BRAINS4

[15] European Commission, *Europass digital credentials infrastructure*. Directorate-General for Employment, Social Affairs and Inclusion, 2022. Available: https://europass.europa.eu/

[16] Diário da República, *Decreto-Lei n.º65/2018 – Regime jurídico dos graus e diplomas do ensino superior*. Lisboa, 2018. Available: https://diariodarepublica.pt/

[17] Direção-Geral do Ensino Superior, *Quadro nacional de qualificações*, 2025. Available: https://www.dges.gov.pt/pt/quadro_qualificacoes

[18] ENIC-NARIC, *The European Recognition Manual for Higher Education Institutions*. Nuffic, 2020. Available: https://www.enic-naric.net/page-EAR-HEI-manual

[19] A. Grech, B. Venkataraman, and F. Miao, *Education and blockchain*. United Nations Educational, Scientific and Cultural Organization, 2022. Available: https://doi.org/10.56059/11599/4131

[20] A. Mikroyannidis, J. Domingue, M. Bachler, and K. Quick, "Blockchain-based descentralised micro-accreditation for lifelong learning," *Interactive Learning Environments*, 2024. Available: https://doi.org/10.1080/10494820.2024.2401485

[21] A. Karale and H. K. Khanuja, "Blockchain technology in education system: A review," *International Journal of Computer Applications*, vol. 178, no. 38, pp. 19-22, 2019. Available: https://doi.org/10.5120/ijca2019919256

[22] D. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," *Decentralized Business Review*, 2008. Available: https://bitcoin.org/bitcoin.pdf

[23] M. Pilkington, "Blockchain technology: Principles and applications," in *Research Handbook on Digital Transformations*, 2016. Available: https://ssrn.com/abstract=2662660

[24] D. Tapscott and A. Tapscott, Blockchain revolution: *How the technology behind Bitcoin and other cryptocurrencies is changing the world*. Portfolio, 2018.

[25] Z. Zeng, S. Xie, H. Dai, X. Chen, and H. Wang, "Blockchain challenges and opportunities: A survey," *International Journal of Web and Grid Services (IJWGS)*, vol. 14, no. 4, 2018. Available: https://www.inderscience.com/offers.php?id=95647

[26] V. Buterin, "A next-generation smart contract and descentralized application platform," 2014. Available: https://ethereum.org/en/whitepaper/

[27] A. Santhi, and P. Muthuswamy, "Influence of blockchain texhnology in manufacturing supply chain and logistics," *Logistics*, vol. 6, no. 1, p.15, 2022. Available: https://doi.org/10.3390/logistics6010015

[28] M. Antonopoulos and G. Wood, *Building smart contracts and DApps*. O'Reilly Media, 2018. Available: https://ethereumbook.info/

[29] K. Christidis and M. Devestsikiotis, "Blockchains and smart constracts for the Internet of Things," *IEEE Access*, vol.4, pp. 2292-2303, 2016. Available: https://doi.org/10.1109/ACCESS.2016.2566339

[30] D. Mingxiao, M. Xiaofeng, Z. Zhe, W. Xiangwei, and C. Qijun, "A review on consensus algorithm of blockchain," in *Proc. IEEE Int. Conf. on Systems, Man, and Cybernetics (SMC)*, 2017, pp. 2567-2572. Available: https://doi.org/10.1109/SMC.2017.8123011

[31] F. Casino, T. Dasaklis, and C. Patsakis, "A systematic litearture review of blockchain-based applications: Current status, classification and open issues," *Telematics and Informatics*, vol. 36, pp. 55-81, 2019. Available: https://doi.org/10.1016/j.tele.2018.11.006

[32] P. Williams, "Does competency-based education with blockchain signal a new mission for universities?" *Journal of Higher Education Policy and Management*, vol.41, no.1, pp. 104-117, 2018. Available: https://doi.org/10.1080/1360080X.2018.1520491

[33] A. Santhi, and P. Muthuswamy, "Influence of blockchain texhnology in manufacturing supply chain and logistics," *Logistics*, vol. 6, no. 1, p.15, 2022. Available: https://doi.org/10.3390/logistics6010015

[34] G. Laatikainen, M. Li, and P. Abrahamsson, "A system-based view blockchain governance". *Information and Software Technology,* vol.157, 107149, 2023. Available: 10.1016/j.infsof.2023.107149

[35] R. Dhillon and P. Sivabalan, "Exploring dimensions of governance for different types of blockchain systems", *The British Accounting Review*, 101588, 2025. Available: https://doi.org/10.1016/j.bar.2025.101588

[36] M. Ibrahimy, A. Norta, and P. Normak, "Bockchain-based governance models supporting corruption-transparency: A systematic literature review", *Blockchain: Research and Applications*, vol. 5, no. 2, 2024. Available: https://doi.org/10.1016/j.bcra.2023.100186

Optimization of BERT for Aspect-Based Sentiment Analysis in Reviews

1st Belén María Ramírez Gabardino
Grupo de Ingeniería de Medios (GIM)
Universidad de Extremadura
Cáceres. Spain belramirez@unex.es

2nd Víctor Gonzalez Morales
Grupo de Ingeniería de Medios (GIM)
Universidad de Extremadura
Cáceres. Spain victorgomo@unex.es

3rd Fernando Broncano Morgado
Grupo de Ingeniería de Medios (GIM)
Universidad de Extremadura
Cáceres. Spain fbroncano@unex.es

4th Mar Avila Vegas
Grupo de Ingeniería de Medios (GIM)
Universidad de Extremadura
Cáceres. Spain
mmavila@unex.es

***Abstract*—In recent years, the analysis of emotions and sentiments in written texts has gained increasing relevance, especially in human communication and interaction. Artificial intelligence has facilitated the automation of this process, addressing the challenge of interpreting emotional nuances that can be subjective even for humans. In this context, Transformer-based models, such as BERT (Bidirectional Encoder Representations from Transformers), have demonstrated outstanding capabilities in processing and understanding texts in English. These models enable precise identification and classification of emotions, enhancing interaction in applications related to sentiment analysis. Additionally, aspect-based sentiment analysis (ABSA) has emerged as a key technique, allowing for a detailed evaluation of emotions associated with different components of a text. This study focuses on optimizing BERT for its application in emotion analysis in written texts, evaluating its performance in various contexts.**

***Index Terms*—Natural Language Processing (NLP), aspect-based sentiment analysis, ABSA, BERT, emotion classification.**

I. Introduction

Sentiment analysis has become a fundamental area in the field of Natural Language Processing (NLP), enabling the identification and understanding of emotions and opinions expressed in text [1]. This study focuses on the exploration and development of advanced techniques for sentiment analysis, utilizing Transformer-based models, particularly BERT (Bidirectional Encoder Representations from Transformers) [2], an architecture that has proven highly effective in complex NLP tasks.

As NLP has evolved, the need to address more complex challenges, such as emotion and opinion classification in textual data, has become evident. In this context, tools like BERT play a crucial role, as their ability to model long-range dependencies and process large volumes of textual data efficiently makes them a powerful solution for tackling these challenges.

Natural Language Processing (NLP) is a field of artificial intelligence aimed at enabling computers to understand, interpret, and generate text in a manner similar to human cognition. This encompasses various tasks such as inference, question answering, and semantic equivalence, among others. Within this domain, sentiment analysis emerges as a key subdiscipline focused on identifying and extracting the emotions and opinions conveyed in textual content. In this context, advanced NLP techniques play a fundamental role in enabling machines not only to understand language but also to capture the emotional nuances it conveys.

The evaluation of sentiment analysis determines the emotional tone of a given text, identifying whether it conveys a positive, negative, or neutral stance. Beyond this general classification, it also enables the detection of specific emotions such as happiness, sadness, anger, or surprise. These capabilities are particularly valuable in domains such as social media monitoring, market analysis, and customer experience optimization, as they provide crucial insights into users' perceptions and emotions regarding products, services, or events [3].

Before examining the techniques and models used for sentiment analysis, it is essential to establish a clear distinction between the concepts of emotions and sentiments. Although these terms are often used interchangeably, each has distinct characteristics that influence how they are interpreted and processed in the context of Natural Language Processing.

A. Difference Between Sentiments and Emotions

Although the terms "sentiments" and "emotions" are frequently used interchangeably, a clear distinction exists in both psychological and linguistic contexts. Emotions are immediate and automatic reactions to a stimulus, characterized by their intensity and short duration [4]. According to Plutchik's theory [5], basic emotions such as anger, fear, surprise, sadness, joy, disgust, and anticipation form the foundation of human emotional experience. These emotions can combine to generate complex emotions, such as love or despair.

In contrast, sentiments are more conscious and enduring interpretations of emotions. While emotions are rapidly triggered in response to an event, sentiments involve more reflective

evaluations that develop over time. Although typically less intense, sentiments tend to be more sustained and long-lasting.

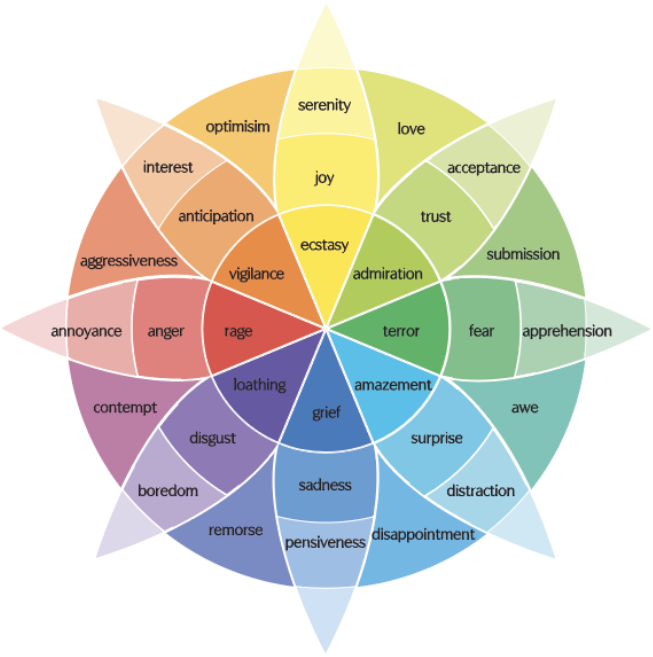


Fig. 1. Plutchik's Wheel of Emotions

Plutchik's Wheel of Emotions [5] is a visual representation that classifies emotions based on their intensity and the relationships between them (figure 1). This model facilitates the understanding of how different emotions and sentiments interact. In the field of NLP, it has been used as a key tool for accurately identifying and labeling emotions in textual data. Its theoretical structure provides a solid foundation for developing more precise and effective sentiment analysis models.

Moreover, language plays a fundamental role in how emotions are expressed and perceived. Through message construction and word selection, emotions can be intensified, softened, or even made ambiguous. As a result, linguistic choices not only reflect an emotion but also profoundly influence how that emotion is experienced by the recipient [6] [7]. This interaction between language and emotion is central to NLP models, which must be capable of capturing linguistic richness to perform accurate sentiment analysis.

Several studies have explored the application of advanced models to improve aspect-based sentiment analysis (ABSA). In this regard, Mughal, Mujtava, Shaikh, Kumar, and Daudpota [8] conducted a comparative analysis between deep neural networks and large language models (LLMs) such as DeBERTa, PaLM, and GPT-3.5-Turbo, highlighting that DeBERTa delivers consistent performance across various ABSA tasks, while PaLM demonstrates competitiveness, particularly in aspect-term and aspect-sentiment analysis (ATSA). Additionally, they noted the domain sensitivity of these models, emphasizing the need to tailor ABSA approaches based on the data type. Similarly, Ansar, Goswami, Chakrabarti, and Chakrobory [9] proposed an efficient methodology for ABSA using BERT through refined aspect extraction. This approach optimizes the process by significantly reducing sentence length and textual complexity. Both studies demonstrate how the use of advanced models such as BERT and PaLM can enhance efficiency and effectiveness in sentiment and aspect-based analysis, opening new avenues for research and applications in the field of Natural Language Processing.

In the context of the previously mentioned advances, this study aims to address key aspects to improve emotion analysis in textual data, particularly in English. To this end, fundamental factors that directly impact the accuracy and effectiveness of emotion classification models are explored. The contributions of this paper are outlined as follows:

1. The study of the impact of data balancing on emotion classification in English, evaluating how class distribution affects model performance, particularly in the detection of underrepresented emotions.
2. The comparison of different hyperparameter configurations, including variations in learning rate, regularization, and schedulers, to determine their influence on model accuracy in emotion classification in English.

II. Methodology

BERT is a deep learning model pretrained on the Transformer architecture that employs a bidirectional approach to process text, allowing it to interpret words based on their full context. Unlike unidirectional models, BERT considers both previous and subsequent words in a sentence, enhancing semantic understanding. Its training is based on two key tasks: the Masked Language Model (MLM), where it predicts hidden words in a sentence, and Next Sentence Prediction (NSP), which allows it to learn relationships between sentences. These features make BERT an effective tool for various natural language processing (NLP) tasks, such as text classification and sentiment analysis.

In this study, we explore the use of BERT for emotion classification in English, employing Google's pre-trained *bert-base-uncased* model, optimized for processing text without distinguishing between uppercase and lowercase letters. The implementation and experimentation were conducted in Google Colab, a platform that facilitates the execution of machine learning models without requiring additional setup, providing access to GPUs and TPUs to accelerate training.

ABSA is an NLP technique that enables the classification of opinions based on specific attributes of an object or entity [10] [1]. Unlike general sentiment analysis, which determines the overall polarity of a text, ABSA breaks down opinions by specific features, providing a more detailed analysis. This methodology is particularly useful in product reviews, where consumers may express differentiated evaluations regarding various aspects of the same item.

To evaluate the impact of data distribution on model performance, two versions were trained under the Aspect-Based Sentiment Analysis (ABSA) approach: one with a balanced dataset and another with an unbalanced dataset. Both models classify reviews into three polarity categories - positive,

negative, and neutral-, allowing for a more precise analysis of how opinions are expressed regarding different aspects of the book. This comparison provides a more detailed view of the model's representativeness and its ability to capture nuances in readers' perceptions.

III. Development

Based on the described methodology, the development of this experiment was structured into several phases, ranging from data preparation to model evaluation.

A. Dataset Preparation

The selected ABSA dataset [11] is an XML file containing book reviews, where each review consists of one or more sentences.

Each sentence is associated with a unique identifier and contains a text representing an opinion about the book. The opinions include the attributes *category* (type of opinion), *polarity* (polarity of the opinion), and *target* (subject of the opinion). This structure facilitates a detailed sentiment analysis, enabling a more precise evaluation of opinions about specific aspects of books.

In order for the data to be properly adapted to the aspect-based sentiment analysis (ABSA), several key actions were taken in its processing and transformation.

B. Data Extraction and Conversion to Tabular Format

The extraction of opinions was carried out using the *xml.etree.ElementTree* library, which efficiently handles XML files in Python. This tool facilitated access to relevant elements within each review, such as the opinion text, the evaluated aspect category, the assigned polarity, and the opinion's subject. Each extracted opinion was stored as a dictionary with the corresponding attributes. Subsequently, the extracted data was saved in a CSV file, enabling its reuse in future analysis phases. If the CSV file already existed, it was loaded directly, avoiding repetition of the extraction process and improving data handling efficiency.

Once the information was extracted, it was converted into a suitable format for analysis. The extracted data was organized into a *DataFrame* using the Pandas library, structuring the information in a tabular format with columns for the previously described attributes. This conversion allowed for easier and more efficient manipulation of the data during the subsequent analysis.

C. Data Cleaning

In this stage, various operations were performed to improve the quality of the dataset. Null values were detected and removed, especially in key columns such as *text* and *target*, which are essential for the model. Duplicate reviews were also identified and removed to ensure that each opinion was considered only once. Additionally, the text was normalized by removing special characters and unnecessary spaces, which allowed for greater consistency in the analysis and subsequent application of processing techniques.

D. Verification of Polarity Distribution

With the dataset prepared, an exploratory analysis was conducted to visualize the distribution of opinions according to their polarity, represented in a pie chart (figure 2). This analysis revealed a significant imbalance: a large majority of opinions exhibited neutral polarity, in contrast to 30% positive and 10% negative opinions.

To mitigate this effect, two versions of the training dataset were generated: (1) an unbalanced dataset that preserves the original polarity distribution, and (2) a balanced dataset in which the class proportions were adjusted to ensure an equal number of positive, negative, and neutral samples. Balancing was achieved through a combination of oversampling and undersampling techniques.

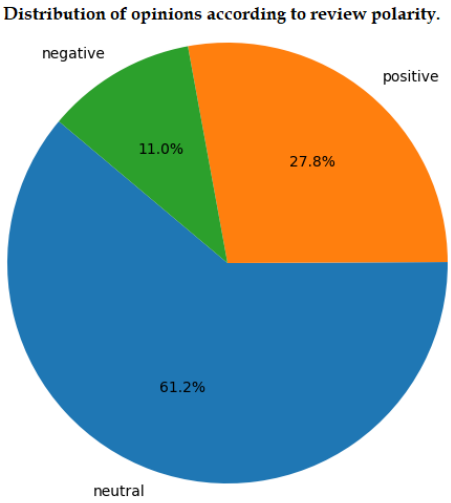


Fig. 2. Distribution of opinions according to review polarity.

This imbalance in the polarity distribution presents a challenge for NLP models, as an imbalanced dataset can affect the accuracy and effectiveness of sentiment analysis. However, this information also facilitates the implementation of data balancing techniques, with the aim of improving model performance in future iterations. Due to this significant class imbalance, two versions of the dataset were used: the original imbalanced dataset, where the polarity distribution remained as it was in the original dataset; and the balanced dataset, in which undersampling techniques were applied to achieve an equitable distribution of polarities.

E. Dataset Splitting

To evaluate the performance of the sentiment analysis model, the dataset was partitioned into two subsets: training and testing. This procedure was carried out for both the original version of the dataset and its balanced version, using the *train_test_split* function from the *sklearn.model_selection* library. An 80% training and 20% testing split was established, ensuring that both subsets contained sufficient examples for learning and model evaluation.

The process began by loading the datasets from their respective locations, followed by the assignment of numerical values to the polarity labels to facilitate processing. Additionally, a combined column was generated, concatenating the text with the polarity label (*text_with_target*), providing additional context during the prediction phase in the Aspect-Based Sentiment Analysis (ABSA) framework.

However, no validation set was generated due to the limited size of the dataset. Since hyperparameter optimization was not the focus at this stage, maximizing the amount of data available for training was prioritized.

F. Tokenization and Data Preparation

The next crucial step in data preparation was tokenization, which transforms raw text into sequences of tokens that can be processed by the BERT model. For this task, the pre-trained BERT tokenizer was used, which operates based on a subword-based scheme. This procedure consists of several stages:

1. Word segmentation: The text is broken down into smaller chunks, allowing infrequent words to be handled by splitting them into subwords.
2. Conversion to indices: Each generated token is assigned a numerical identifier within the BERT vocabulary.
3. Addition of special tokens: Tokens such as [CLS] (start of sequence) and [SEP] (separator) are added, which are essential for BERT to understand the structure of the input.
4. Padding and truncation: Padding or truncation techniques are applied to ensure that all sequences have uniform length, facilitating batch processing during training.

G. Loading the BERT Model

In this study, the pre-trained *BertForSequenceClassification* model was used, designed for text classification tasks and based on the BERT architecture. This architecture uses attention mechanisms to capture contextual relationships within a text sequence.

Since the pre-trained model was not specifically designed for sentiment classification, its final layers were modified to allow for multi-class classification into three polarity categories: positive, neutral, and negative. Throughout this process, hyperparameters were adjusted, and changes were made to the network architecture to optimize its performance in sentiment classification.

To facilitate efficient data handling during training, a custom class called *OpinionDataset* was implemented, which organizes the data into batches and manages attention masks so the model can identify relevant parts of the sequences. Finally, to evaluate the impact of class distribution on model performance, two versions were trained: one with the original imbalanced dataset and one with a balanced dataset.

H. Impact of Data Balance on Model Training

In this scenario, the polarity classes were not represented equitably, which could induce bias in the model toward the majority classes. The AdamW algorithm, an enhanced version of Adam that incorporates L2 regularization through weight decay, was used for the optimization process. This technique helps control the magnitude of the model parameters, reducing the risk of overfitting and improving its generalization ability.

Training was carried out for three epochs, during which the loss function was calculated to measure the discrepancy between the model's predictions and the true labels. Through the backpropagation mechanism, the model's weights were updated to minimize the loss. The use of minibatches optimized training time and stabilized the parameter updates. However, the uneven class distribution in this dataset may have introduced bias, favoring the prediction of more frequent classes at the expense of the less represented ones.

I. Impact of Data Balance on Model Training

To mitigate the effects of class bias, a second version of the model was trained using the balanced dataset, where all polarity classes are represented equitably. This approach aims to provide the model with a more homogeneous distribution of examples, allowing it to learn more effectively and improve its performance in minority classes.

The training procedure remained consistent with that used for the unbalanced dataset, including the use of the AdamW optimizer, the calculation of the loss function, and the backpropagation mechanism. Training was also carried out for three epochs, as the model showed rapid convergence without showing signs of overfitting.

IV. Results

To perform a detailed comparison between the sentiment classification models developed, a classification report was generated that provides a comprehensive evaluation of model performance, displaying both class-specific metrics and general metrics, along with confusion matrices.

A. Evaluation of the Unbalanced Model

This section presents the results () of the aspect-based sentiment analysis model using BERT IV-A on an unbalanced dataset of 767 instances across three categories: positive, neutral, and negative.

Classes	Precision	Recall	F1-score	Instances
Positive	0.78	0.76	0.77	216
Neutral	0.84	0.87	0.86	453
Negative	0.63	0.56	0.59	98
Accuracy			0.80	767
Macro Average	0.75	0.73	0.74	767
Weighted Average	0.80	0.80	0.80	767

TABLE I
CLASSIFICATION REPORT OF THE UNBALANCED MODEL

Class-wise Metric Analysis

The model performs notably well in the neutral class, with a precision of 84% and a recall of 87%. This performance can be attributed to the high representation of this class in the dataset (453 instances), which facilitated the learning of its characteristic patterns. As a result, the model is able to identify neutral instances with high precision and consistency.

On the other hand, the positive class presents a precision of 78%, indicating that the model has a good ability to recognize positive opinions. However, its recall of 76% and F1-score of 0.77 suggest that there are still cases where the model fails to capture all possible positive instances. This could be explained by the distribution of the dataset or the similarity between the positive and neutral classes, which may lead to confusion in classification.

Finally, the negative class shows the lowest performance among the three categories, with a precision of 63% and a recall of 56%. The low representation of this class in the dataset (98 examples) negatively impacts the model's ability to learn distinctive patterns associated with negative opinions. This highlights the need to improve data distribution equity to optimize the model's performance in classifying this category.

Global Metric Analysis

Globally, the model's accuracy is 80%, indicating a good overall performance. However, this value may be influenced by the neutral class, which is the most represented in the dataset. This bias could lead to an overestimation of the model's performance in classifying this class, minimizing the impact of the minority classes. Regarding the macro average, which calculates the arithmetic mean of the metrics without weighting by class size, the model achieves a precision of 75%, a recall of 73%, and an F1-score of 74%. These results suggest an acceptable overall performance, but with disparities in class classification. Finally, the weighted average precision and recall are 80%, reflecting the influence of the neutral class on the model's global performance due to its higher representation in the dataset.

Confusion Matrix Analysis and Interpretation

The confusion matrix presented (figure 3) shows the performance of an aspect-based sentiment analysis (ABSA) model applied to an unbalanced dataset.

The model demonstrated a high performance in classifying neutral opinions, correctly identifying 395 out of 453 instances of this class. However, there was also a tendency to misclassify positive (42) and negative (32) opinions as neutral, indicating a bias towards the majority class. Additionally, 36 neutral opinions were classified as positive and 22 as negative, reflecting confusion in the delineation of these categories.

The confusion between the positive and neutral classes was particularly noticeable. Although the model correctly classified 164 positive instances, 42 were misclassified as neutral, while 36 neutral opinions were confused with positive. This overlap may be due to the inherent subjectivity of certain texts, which may share linguistic characteristics with neutral opinions. Additionally, 10 positive instances were classified as negative and 11 negative as positive, suggesting difficulties in distinguishing the extremes of polarity.

Regarding the negative class, which is the least represented in the dataset, the model correctly identified 55 instances, which is 56% of the opinions labeled as negative. However, 32 negative opinions were classified as neutral and 11 as positive, highlighting the model's difficulty in recognizing distinctive patterns in this category. Furthermore, the presence of 22 neutral and 10 positive opinions misclassified as negative underscores a greater uncertainty in classifying this class.

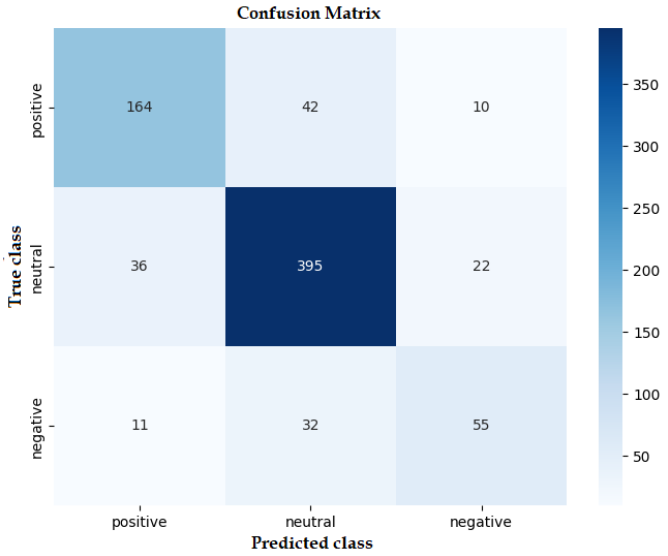


Fig. 3. Confusion Matrix of the Unbalanced Model.

In conclusion, the confusion matrix analysis highlights the limitations of the model in differentiating opinions, especially in the minority classes. The tendency to classify instances as neutral and the confusion between adjacent classes suggest that balancing the data distribution is key to improving model precision and reducing biases in emotion classification.

B. Evaluation of the Balanced Model

The classification report evaluates the performance of an ABSA model on a balanced dataset IV-B, allowing a fairer comparison between the positive, neutral, and negative classes. Unlike the unbalanced dataset, the test set contains 225 instances evenly distributed, providing a more just framework for measuring the model's ability to classify each polarity.

Classes	Precision	Recall	F1-score	Instances
Positive	0.83	0.70	0.76	74
Neutral	0.65	0.68	0.66	75
Negative	0.64	0.70	0.67	76
Accuracy			0.69	225
Macro Average	0.70	0.69	0.70	225
Weighted Average	0.70	0.69	0.70	225

TABLE II
CLASSIFICATION REPORT OF THE BALANCED MODEL

Class-wise Metric Analysis

As shown in table IV-B, the model achieves an accuracy of 83% for the positive class, indicating a good level of accuracy in identifying positive instances with a low number of false positives. Additionally, it presents a recall of 70%, suggesting that the model is capable of recognizing most positive instances, although there are still some that are not correctly

classified. The F1-score is 0.76%, reflecting a good balance between precision and recall, confirming that performance in this class is effective and well-balanced.

For the neutral class, the model achieves a precision of 65%, indicating that a significant proportion of instances from this category are confused with other classes, increasing false positives. The recall is 68%, reflecting good performance in detecting neutral opinions, though slightly lower than the model trained with the unbalanced dataset. This could be due to the even distribution of instances in the dataset, making it more challenging for the model to clearly distinguish this class. The F1-score of 0.66 also indicates a moderate performance in classifying neutral opinions, with some difficulty in differentiating this category from others.

For the negative class, the model achieves a precision of 64%, reflecting a significant improvement over the model trained on the unbalanced dataset. This result suggests that the model has learned to identify negative instances better by removing the bias towards the majority classes. The recall for the negative class is 70%, indicating improved identification of negative instances compared to the unbalanced model. The F1-score of 0.67 confirms that performance in this class is acceptable and that the even distribution of data has favored the model's learning in detecting negative opinions.

In summary, balancing the dataset has allowed a more equitable evaluation of the model, eliminating the bias toward the majority class and improving performance in classifying negative opinions. While performance in the neutral class has decreased compared to the unbalanced dataset, the model presents a more uniform distribution of errors, which allows for a more objective evaluation of its classification performance.

Global Metrics Analysis

At a global level, the model's accuracy is 69%, which represents an adequate performance considering the balanced data. Although this value is lower than that obtained with the unbalanced dataset, it reflects a more realistic evaluation of the model, as it is not influenced by a majority class. Regarding the macro average, which calculates the arithmetic mean of the metrics without weighting by class size, the precision, recall, and F1-score scores are 0.70, indicating equitable performance across classes and suggesting that the balanced data has reduced the bias towards a specific class. Finally, the weighted average is also 0.70 for all metrics, confirming that the model maintains consistent performance across the three classes, without disproportionately favoring any particular one.

Confusion Matrix Analysis and Interpretation

The confusion matrix presented in figure 4 provides a clear visualization of the ABSA classification model's performance on a balanced dataset.

The model showed acceptable performance in classifying positive opinions, correctly identifying 52 instances of this class. However, there was a tendency to confuse neutral (6 instances) and negative (5 instances) opinions with positive ones, suggesting that the model has difficulty distinguishing between positive and other categories. Additionally, 10 false negatives were recorded, where positive opinions were classified as neutral, and 12 as negative. This pattern highlights the model's difficulty in correctly identifying positive opinions, particularly in distinguishing them from the neutral class.

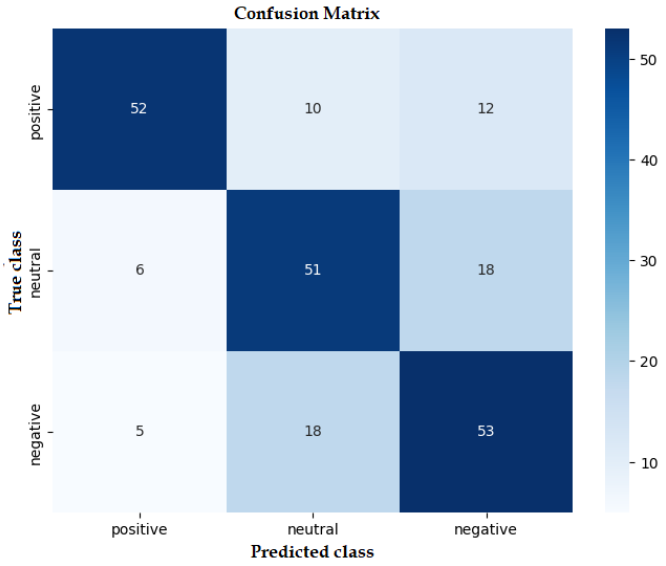


Fig. 4. Confusion matrix of the balanced model.

Regarding neutral opinions, the model showed good performance by correctly classifying 51 instances. However, 10 false positives from the positive class and 18 from the negative class were recorded, indicating that the model tends to incorrectly classify positive and negative opinions as neutral in situations of greater uncertainty. This behavior reflects a conservative approach in classification. Additionally, 6 false negatives were identified, where neutral opinions were classified as positive and 18 as negative. This confusion suggests that the model has difficulty clearly distinguishing neutral opinions from other classes, affecting the overall classification accuracy.

For negative opinions, the model correctly identified 53 instances, indicating adequate performance in this class. However, 12 false positives from the positive class and 18 from the neutral class were found, revealing that the model has difficulty distinguishing negative opinions from the other classes, especially from neutral opinions. Furthermore, 18 false negatives showed negative opinions classified as neutral and 5 as positive. These errors reinforce the model's tendency to confuse the negative class with the neutral one, which could impact the analysis's precision.

The confusion matrix analysis (figure 4) reveals that the model has difficulty distinguishing between the positive and negative classes, which is reflected in the significant presence of false positives and false negatives between these two classes. The neutral class, despite achieving reasonably good performance, tends to absorb errors from the other classes,

Class	Precision (NoB)	Precision (B)	Recall (NoB)	Recall (B)	F1-Score (NoB)	F1-Score (B)
Positive	0.78	0.83	0.76	0.70	0.77	0.76
Neutral	0.84	0.65	0.87	0.68	0.86	0.66
Negative	0.63	0.64	0.56	0.70	0.59	0.67

TABLE III
COMPARISON OF METRICS BETWEEN THE BALANCED MODEL (B) AND THE UNBALANCED MODEL (NOB)

suggesting that the model might be adopting a conservative approach in its classification. This error pattern is most notable in the extreme classes (positive and negative), where the model shows some confusion due to the semantic proximity of opinions.

Balancing the dataset has been crucial for optimizing the model's performance, especially in the detection of negative opinions. Compared to models trained on unbalanced data, a significant improvement was observed in the identification of the negative class. This underscores the importance of using a balanced dataset to minimize biases and ensure more accurate and representative classification across all classes.

C. Comparison of Trained Models

The comparative analysis between models trained with balanced (B) and unbalanced (NoB) datasets, shown in table IV-B, illustrates that using a balanced dataset contributes to improved performance in certain underrepresented classes, particularly the negative class, although it may involve trade-offs in other metrics.

In the case of the positive class, the model trained with balanced data experiences a reduction in precision, suggesting an increase in the detection of instances of this class accompanied by a higher number of false positives. This decrease in precision, coupled with a slight improvement in recall, indicates that the model classifies a larger number of opinions as positive, although not always correctly.

For the neutral class, a slight reduction in F1-score is observed, mainly attributed to a decrease in recall. This suggests that the model trained with balanced data has a lower capacity to correctly identify instances of this class compared to the unbalanced model, which may indicate overfitting affecting its performance in this category.

On the other hand, using a balanced dataset has proven particularly beneficial for classifying the negative class. There is an improvement in both precision and recall, resulting in a significant increase in F1-score. These results suggest that the balanced model has improved its ability to identify negative opinions, a crucial advancement considering the initial underrepresentation of this category in the original dataset.

V. CONCLUSIONS

One of the main challenges observed in the results is the model's difficulty in correctly identifying the neutral class, which corresponds to phrases without a clearly positive or negative polarity. In most predictions, the model tends to classify texts as either positive or negative, failing to properly recognize intermediate expressions. This behavior reveals limitations in processing inherently neutral texts, possibly due to the training methodology or the inherent complexity of classifying across multiple sentiment categories. It highlights a constraint in the model's generalization capability when dealing with such inputs.

On the other hand, the use of a balanced dataset has proven beneficial for classifying the negative class, enabling the model to capture a greater diversity of opinions in sentiment analysis. However, the impact of data balancing on the positive and neutral classes suggests the presence of adverse effects, particularly a decrease in performance when classifying neutral opinions, possibly due to slight overfitting. These results highlight the importance of implementing more advanced techniques, such as hyperparameter tuning or synthetic data generation, to enhance the representativeness of the training set and mitigate potential model limitations.

Furthermore, the absence of a validation phase during training, due to the limited size of the dataset, has limited the optimization capacity of the model. To enhance the model's generalization capability and reliability, it is recommended to expand the dataset and implement a robust validation phase, coupled with hyperparameter fine-tuning to mitigate overfitting.

ACKNOWLEDGMENTS

This initiative is carried out within the framework of the funds of the Recovery, Transformation and Resilience Plan, financed by the European Union (Next Generation) - National Cybersecurity Institute (INCIBE) in the project C109/23 "Strategic Project UEx (Polytechnic School of Cáceres) - INCIBE".

REFERENCES

[1] B. Liu, *Sentiment Analysis and Opinion Mining*. San Rafael, Calif.: Morgan & Claypool, 2012.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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: Human Language Technologies*. Association for Computational Linguistics, 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[3] E. Cambria, S. Poria, A. Gelbukh, and B. Liu, "Sentiment analysis is a big suitcase," in *Computational Intelligence and Data Science*. Springer, 2017, pp. 1–10.

[4] K. R. Scherer and A. Moors, "Emotion: Theories and models," in *The Social Psychology of Emotion*, J. P. Forgas, K. D. Williams, and W. von Hippel, Eds. Psychology Press, 2019, pp. 1–23.

[5] R. Plutchik, "A general psychoevolutionary theory of emotion," in *Emotion: Theory, Research, and Experience*, R. Plutchik and H. Kellerman, Eds. New York: Academic Press, 1980, vol. 1, pp. 3–33.

[6] G. Lakoff and M. Johnson, *Metaphors We Live By*. Chicago: University of Chicago Press, 1980.

[7] H. H. Clark, *Using Language*. Cambridge: Cambridge University Press, 1996.

[8] N. Mughal, G. Mujtaba, S. Shaikh, A. Kumar, and S. M. Daudpota, "Comparative analysis of deep natural networks and large language models for aspect-based sentiment analysis," *IEEE Access*, vol. 10, p. 3386969, 2024.

[9] W. Ansar, S. Goswami, A. Chakrabarti, and B. Chakrobory, "An efficient methodology for aspect-based sentiment analysis using bert through refined aspect extraction," *Journal of Intelligent & Fuzzy Systems*, vol. 40, no. 5, pp. 9627–9644, 2021.

[10] Y. C. Hua, P. Denny, K. Taskova, and J. Wicker, "A systematic review of aspect-based sentiment analysis: Domains, methods, and trends," *Artificial Intelligence Review*, vol. 57, p. 296, 2024.

[11] T. Álvarez López, M. Fernández-Gavilanes, E. Costa-Montenegro, J. Juncal-Martínez, S. García-Méndez, and P. Bellot, "A book reviews dataset for aspect based sentiment analysis," in *Proceedings of the 8th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2017)*, 2017.

Enhancing Word-Level Adversarial Attack Generation Using Large Language Models

1st Natalia Madrueño
Data Science Laboratory
Rey Juan Carlos University
Móstoles (Madrid), Spain
natalia.madrueno@urjc.es

2nd Alberto Fernández-Isabel
Data Science Laboratory
Rey Juan Carlos University
Móstoles (Madrid), Spain
alberto.fernandez.isabel@urjc.es

3rd Andrés Caro
Media Engineering Group
University of Extremadura
Cáceres, Spain
andresc@unex.es

4th Isaac Martín de Diego
Data Science Laboratory
Rey Juan Carlos University
Móstoles (Madrid), Spain
isaac.martin@urjc.es

***Abstract*—Recent advances in Natural Language Processing (NLP) rely on black-box models that provide predictions with confidence scores but lack transparency in their decision-making. This opacity complicates the identification of model vulnerabilities. A key strategy for analyzing these weaknesses is the generation of text adversarial examples, where subtle word-level modifications can lead models to incorrect predictions while preserving semantic meaning. This paper proposes a novel method for generating adversarial examples using Large Language Models (LLMs) to perturb critical words that significantly impact predictions. The approach employs LLM-driven instruction prompts to replace vulnerable words with synonyms or insert neutral words adjacent to them. Experiments on sentiment classification tasks demonstrate the method's effectiveness in misleading models while maintaining text coherence. The results highlight the advantages of the proposed approach over existing LLM-based adversarial attack strategies.**

***Index Terms*—Adversarial Perturbation, LLM-based Attacks, Semantic-preserving Perturbations, Word-level Perturbation**

I. Introduction

The latest advancements in Artificial Intelligence (AI) have been driven by adopting highly complex models, deep neural networks in particular [1]. Despite their remarkable performance across diverse tasks, these models are often opaque black boxes, offering limited or no interpretability of their internal mechanisms [2].

The lack of transparency in black-box models extends beyond challenges in explainability. It can compromise the robustness of predictions and hinder the identification of model vulnerabilities, leading to potential ethical and legal concerns [3]. One effective technique for analyzing these vulnerabilities is the generation of adversarial examples—subtly modified inputs designed to mislead models while preserving their semantic integrity [4]. By studying adversarial examples, researchers can enhance model robustness and develop mitigation strategies against such attacks [5].

In the domain of Natural Language Processing (NLP), black-box models frequently rely on predictions without exposing their underlying decision-making processes [6]. This opacity complicates the interpretability of these models and makes it particularly challenging to design adversarial attacks that effectively exploit their weaknesses.

Without insight into how these models weigh different linguistic features, researchers must develop attack strategies that rely on indirect feedback and empirical observations. While adversarial text perturbations can occur at multiple levels—such as character, word, and sentence—word-level modifications have proven especially effective in altering model predictions. This is because they introduce subtle but meaningful changes that can significantly impact classification outcomes while maintaining the original text's readability, fluency, and coherence.

Additionally, word-level perturbations are less likely to be detected by automated defense mechanisms than character-level noise, making them a particularly robust and practical approach for adversarial attacks in NLP.

This paper introduces a novel adversarial attack methodology focusing exclusively on word-level perturbations using Large Language Models (LLMs). The proposed approach employs LLM-driven instruction prompts to modify critical words that significantly impact the model's decision. Specifically, words identified as highly influential in the prediction process are perturbed by replacing them with synonyms. Moreover, some innocuous words are also included in the text to produce additional perturbations. These perturbations subtly alter the input while preserving its original meaning, increasing the likelihood of misclassifications by the target model.

To evaluate the effectiveness of the proposed approach, experiments were conducted on a well-known sentiment classification dataset: single-sentence reviews. Various black-box victim models were tested to assess the performance of the attack across different real-world scenarios.

The rest of this paper is structured as follows. Section II reviews related work. Section III details the proposed word-level adversarial attack method. Section IV presents the experimental setup and results, and Section V concludes the paper and provides future guidelines.

II. Related work

This section reviews previous research on adversarial attacks and their different types.

Adversarial attacks are designed to manipulate victim models by introducing carefully crafted modifications to input data that remain imperceptible to human observers [4]. These modified inputs, known as adversarial examples, introduce subtle perturbations that can mislead models into making

incorrect or unintended predictions [7]. Due to their effectiveness, adversarial examples have gained significant attention across various fields of AI.

These attacks can be categorized based on the level of knowledge an attacker has about the victim model. The two main types are white-box and black-box attacks [8].

In white-box attacks, attackers have full access to the model's internal architecture, parameters, and gradients, allowing them to craft highly effective perturbations [9]. Within NLP, gradient-based white-box attacks [10] have been widely explored, leveraging the inner gradients of victim models to identify the most influential words or tokens and apply targeted modifications [11].

In contrast, black-box attacks operate without direct access to the victim model's internal workings [2]. These attacks are particularly challenging as they must infer model vulnerabilities based on limited observable output.

Black-box attacks can be divided into three subcategories based on the level of output information available to the attacker [11]: Score-based attacks, Decision-based attacks, and Blind attacks.

In the first type of attack, the attacker has access to confidence scores or probability distributions over the model's output classes, allowing them to modify the input based on this feedback [12].

In the second type of attack, the attacker can only observe the final classification decision (e.g., "positive" or "negative") without additional score information, making attack optimization more difficult [13].

Finally, in the third type of attack, the attacker has no information about the model's output, requiring them to rely on general linguistic principles or heuristic strategies to generate adversarial examples [14].

Focusing on the proposal, it addresses the generation of adversarial examples for score-based black-box attacks. By leveraging model confidence scores, the proposed approach systematically applies text perturbations that maximize the likelihood of misleading the victim model. This method allows for more efficient adversarial attacks while maintaining semantic coherence, making it a practical tool for evaluating the robustness of NLP models.

III. Proposal

This section presents the key components of the proposed method for generating text adversarial attacks using decoder-based LLMs. The core strength of the proposed approach lies in the orchestration of different LLM instruction prompts to introduce word-level perturbations. These prompts strategically modify the vulnerable words within an input text to generate adversarial examples. As a result, the method effectively facilitates the deception of targeted victim models while preserving the same original semantic meaning of an original input text.

The method is decomposed into two steps. First, the most vulnerable words in the input text are identified and ranked by calculating a Word Importance Ranking (WIR) [12]. Second, word-level perturbations are applied to these vulnerable words by generating their corresponding LLM synonyms and LLM word insertions. Figure 1 provides an overview of this entire process of generating text adversarial examples. The following sections give a detailed explanation of each step.

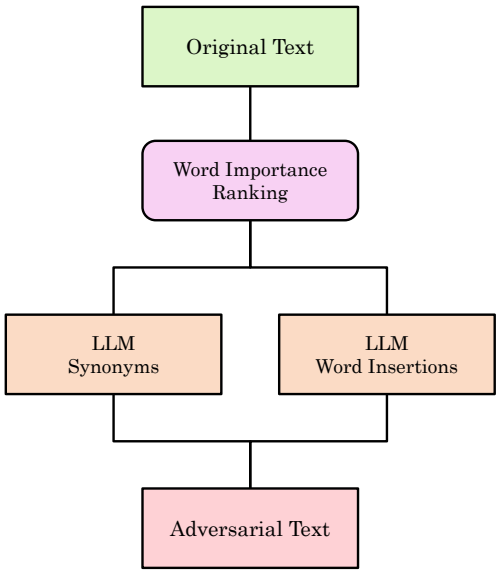


Fig. 1. Proposed method for generating adversarial examples using word-level LLM perturbations. First, the most vulnerable words from an input text are identified and selected by computing a WIR. Then, these vulnerable words are perturbed by replacing them with either their LLM synonyms or by introducing innocuous LLM word insertions to their left or right.

A. Word Importance Ranking (WIR)

Delving into the proposed adversarial example generation method, it selects the n most vulnerable words of an original input text. These vulnerable words are detected and ranked by computing a WIR, which assesses the potential influence of each word on the prediction of a targeted victim model.

For this purpose, the input text X is first tokenized to extract the words that constitute it $W = \{w_1, w_2, \ldots, w_m\}$. The importance of each word $w_i \in W$ is then measured by assessing the impact of its omission on the predictions generated by the victim models. This is done by removing the word w_i in the original input text X and measuring how this removal changes the confidence or probability scores of the victim model. Words whose omissions lead to more significant changes in the output probability of confidence scores of a victim model are considered to be more influential.

By calculating the degree of influence for each $w_i \in W$, a WIR is established as $WIR = (v_1, v_2, ...v_n)$, where $v_j \in W$. The most vulnerable words from an input text can be identified and selected by filtering the n words at the top of this ranking.

B. Word-level perturbations

Once the n most vulnerable words have been selected, the proposed method introduces word-level modifications to the original input text. Specifically, the n most vulnerable words and their immediate surrounding context are systematically

perturbed to fool victim models. For each vulnerable word v_j, and in the order established by the WIR, a series of candidate word-level perturbations are iteratively generated and evaluated.

Specifically, two types of word-level perturbations are explored using LLM instruction prompts: LLM synonyms and LLM word insertions. In the LLM synonyms, a vulnerable word v_j is replaced with one of its contextually appropriate synonyms. In contrast, in the LLM word insertions, extra neutral words are added either to the left or the right of the vulnerable word v_j, subtly altering the text without changing its overall meaning.

Thus, for each vulnerable word v_j and in the order determined by WIR, the corresponding sets of LLM synonyms $S_j = \{s_{j1}, s_{j2}, \ldots, s_{ja}\}$, LLM left word insertions $L_j = \{l_{j1}, l_{j2}, \ldots, l_{jb}\}$ and LLM right word insertions $R_j = \{r_{j1}, r_{j2}, \ldots, r_{jc}\}$ are generated and evaluated. Subsequently, the set of LLM word-level perturbations to be attempted for the word v_j is defined as $P_j = S_j \cup L_j \cup R_j = \{s_{j1}, \ldots, s_{ja}, l_{j1}, \ldots, l_{jb}, r_{j1}, \ldots, r_{jc}\} = \{p_{j1}, \ldots, p_{jd}\}$.

If one of these LLM word-level perturbations $p_{jd} \in P_j$ cause the victim model to change its prediction with sufficient confidence, the first perturbation that does so is applied, and the resulting input text modified with this perturbation is returned as the generated adversarial example X_{adv}. Otherwise, the modification that provokes the most significant effect in changing the predictions of the victim model is chosen to perturb the word v_j, and the algorithm keeps iterating over the next word in WIR. It is important to note that, for each vulnerable word v_j, either a LLM synonym or LLM word insertion is applied, not both simultaneously.

The pseudo code for this complete text adversarial generation proposal can be seen in Algorithm 1. It encompasses both the computation of the WIR and the application of word-level perturbations.

Algorithm 1: Text Adversarial Example Generation

Input : Original input text X
Output: Generated text adversarial example X_{adv}

$W \leftarrow ExtractWords(X)$;
$X_{adv} \leftarrow X$;
for $v_j \in WIR(W)$ **do**
  $P_j \leftarrow GenerateSynonymsInsertions(v_j, X_{adv})$;
  for $p_{jd} \in P_j$ **do**
    if $ModelIsDeceived(p_{jd}, X_{adv})$ **then**
      return $PerturbWord(p_{jd}, X_{adv})$;
    else if $DeceptionIsIncreased(p_{jd}, X_{adv})$ **then**
      $X_{adv} \leftarrow PerturbWord(p_{jd}, X_{adv})$;
return X_{adv}

IV. Experiments

This section details the experiments conducted to assess the quality and effectiveness of the proposed text adversarial example generation method. The evaluation is focused on two critical aspects of text adversarial examples: the ability to change model predictions (Attack Success Rate (ASR)) and the capacity to preserve the semantic meaning of the original input text (semantic preservation). Comparisons are made against representative state-of-the-art approaches and multiple targeted victim models.

Prompt Sentiment Analysis
Determine whether the sentiment of the following sentence is positive or negative. Answer only with the word "Positive" or "Negative". Sentence: "{}" Answer:

TABLE I
LLM INSTRUCTION PROMPT USED TO CLASSIFY THE SENTIMENT OF THE SST-2 SENTENCES. THIS INSTRUCTION PROMPT IS EMPLOYED BY THE INSTRUCT VERSIONS OF THE TARGETED VICTIM MODELS GEMMA 2 9B, LLAMA 3.1 8B, QWEN 2.5 7B AND YI 1.5 6B.

The Binary Stanford Sentiment Treebank (SST-2) dataset has been utilized as the benchmark for evaluating different text adversarial attacks. This dataset encompasses a binary sentiment classification on sentence-level movie reviews from Rotten Tomatoes [1], where the sentiment of these sentences is labeled as either positive or negative. Specifically, the sentences from the original validation split of 872 examples were utilized as the original input sentences from which adversarial attacks to targeted victim models are attempted.

Four open-weight and decoder-based LLMs have been selected as the targeted models to evaluate the effectiveness of adversarial attacks against recent victim models. In particular, the instruct versions of Gemma 2 9B [15], Llama 3.1 8B [16], Qwen 2.5 7B [17], and Yi 1.5 6B [18] models have been chosen as the targeted victim models to attack. These models achieve a binary classification of input sentences according to their positive or negative sentiment. Specifically, this classification is carried out through a zero-shot manner, using the instruction prompt defined in Table I. For each input sentence, it is assumed that the victim models only output the final predicted sentiment label (positive or negative) and its probability. Thus, this setup reflects a realistic black-box scenario where there is no access to or knowledge of the inner workings of the model.

Text adversarial examples are generated using the closed-source and decoder-based LLM Generative Pre-Trained Transformer 4 Omni (GPT-4o) mini model. Thanks to its extraordinary text generation capabilities, context-aware word-level perturbations can be introduced into the original SST-2 sentences to deceive targeted victim models. In particular, two types of word-level perturbations are produced by using GPT-4o mini: LLM synonyms and LLM word insertions. Both LLM-generated perturbations are produced in a zero-shot manner, using the instruction prompts defined in Table II.

The state-of-the-art baseline methods analyzed for comparison against the proposal consist of single LLM word-level

[1] https://www.rottentomatoes.com/

Perturbation	Prompt
Synonyms	Generate a list of synonyms for the target word in the context of the sentence below. Limit to bullet points for each suggested synonym. Sentence: "{}" Word: "{}" Answer: -
Word Insertions	Generate a list of neutral words that could naturally be inserted at the position marked by [INSERTION] in the sentence below. Limit to bullet points for each suggested insertion. Sentence: "{}" Answer: -

TABLE II
LLM INSTRUCTION PROMPTS FOR GENERATING DIFFERENT WORD-LEVEL PERTURBATIONS ON THE ORIGINAL SENTENCES FROM THE SST-2 DATASET. THESE INSTRUCTION PROMPTS ARE LATER USED BY GPT-4O MINI TO GENERATE TEXT ADVERSARIAL EXAMPLES.

SST-2 Adversarial Attack	ASR			
	Gemma	Llama	Qwen	Yi
LLM Synonyms	0.60	0.64	0.62	0.65
LLM Word Insertions	0.57	0.66	0.61	0.63
Proposed method (LLM Synonyms + LLM Word Insertions)	0.77	0.80	0.80	0.80

TABLE III
ASR FOR THE EVALUATED ADVERSARIAL ATTACKS PERFORMED ON THE SST-2 DATASET. GEMMA, LLAMA, QWEN, AND YI ARE THE INSTRUCT VERSIONS OF THE GEMMA 2 9B, LLAMA 3.1 8B, QWEN 2.5 7B, AND YI 1.5 6B LLM MODELS, RESPECTIVELY. LLM SYNONYMS AND LLM WORD INSERTIONS REFER TO POPULAR LLM STATE-OF-THE-ART ADVERSARIAL ATTACKS.

Prompt Semantic Similarity
Determine whether the following two sentences are semantically similar. Answer "YES" if they are semantically similar, or "NO" otherwise. Sentence 1: "{}" Sentence 2: "{}" Answer:

TABLE IV
LLM INSTRUCTION PROMPT EMPLOYED TO DETERMINE WHETHER THE ADVERSARIAL EXAMPLES GENERATED ARE SEMANTICALLY SIMILAR TO THE ORIGINAL INPUT TEXTS OF THE SST-2 DATASET. THIS PROMPT IS LATED USED BY GPT-4O.

perturbation strategies, which are widely adopted for generating text adversarial examples. Specifically, two frequent LLM state-of-the-art approaches are evaluated: either the application of only LLM synonyms or the usage of only LLM word insertions [12], [14]. Note that, in contrast to these LLM state-of-the-art strategies, the proposed method combines both LLM synonyms and LLM word insertions within the same sentence, enabling a potentially more effective adversarial attack.

A maximum of the top 40% most vulnerable words are perturbed in the selected state-of-the-art baselines and the proposed adversarial attack strategy. Accordingly, words are perturbed in the order established by the WIR either until a successful adversarial example is generated or until this maximum percentage is reached, stopping the attack process when this limit is met.

ASR [7], [12] has been calculated to evaluate the effectiveness of the proposed method against the previous state-of-the-art adversarial attack approaches. This metric quantifies the proportion of text adversarial examples that successfully fool a targeted victim model into producing incorrect final predictions.

Table III presents the results of computing ASR for both the evaluated state-of-the-art baselines and the proposed adversarial attack method on the SST-2 dataset. As it can be seen, it is evident that the proposed adversarial example generation approach, which combines both LLM synonyms and LLM word insertions, significantly outperforms the sole usage of either LLM synonyms or LLM word insertions alone. Specifically, the proposed approach achieves ASR values of up to 0.80, with improvements reaching up to 20 percentage points over the weakest baseline comparison. The enhancement in the ASR of the proposal over the evaluated baselines is consistent across all the targeted victim models analyzed, with a notable difference in ASR observed in all of them.

Nevertheless, when evaluating the quality of an adversarial attack technique, it is crucial to measure its ability to alter the predictions of a victim model but also to examine how well its generated adversarial examples preserve the semantic meaning of the original input text. Therefore, the degree of the semantic preservation of the analyzed state-of-the-art approaches and the proposed attack method, is also measured and compared.

In particular, the semantic preservation of adversarial attacks is assessed by employing GPT-4o in a zero-shot manner. To this end, the instruction prompt defined in Table IV was utilized to ascertain how the adversarial texts produced by adversarial attack strategies are semantically similar to their corresponding original input sentences. The degree of semantic preservation of an adversarial attack technique is thus measured by calculating the percentage of semantically similar pairs between the original input texts and their corresponding adversarial examples.

Table V shows the degree of semantic preservation of the analyzed state-of-the-art baselines and the proposed adversarial example generation method on the SST-2 dataset. As it can be observed, and according to GPT-4o, all the evaluated adversarial attacks can preserve the original semantic meaning of the input texts to a high degree. While the LLM synonyms approach may be slightly better at preserving the original semantics, the difference with LLM word insertions and the proposed adversarial attack method is relatively small - especially when compared to the significantly higher ASR achieved by the proposed method. In particular, the most significant disparity in semantic preservation between the proposed method and the best semantic-preserving baseline is no more significant than 6 percentage points.

Consequently, it can be concluded that the proposed adversarial example generation method of combining LLM synonyms and LLM word insertions can effectively attack victim models. Compared to other popular state-of-the-art methods, it has been demonstrated to surpass significantly the ASR rate in all the evaluated decoder-based LLMs victim models while preserving the semantic meaning of the original sentence to a high degree.

SST-2 Semantic Preservation	Semantically Similar			
	Gemma	Llama	Qwen	Yi
LLM Synonyms	0.95	0.95	0.95	0.96
LLM Word Insertions	0.87	0.90	0.90	0.92
Proposed method (LLM Synonyms + LLM Word Insertions)	0.89	0.90	0.91	0.93

TABLE V
PROPORTION OF TEXT ADVERSARIAL EXAMPLES THAT PRESERVE THE ORIGINAL SEMANTIC MEANING OF THE SST-2 DATASET, ACCORDING TO GPT-4O.

V. CONCLUSIONS

This paper introduces a novel method for generating text adversarial examples focused on word-level perturbations to attack targeted victim models. The approach leverages LLMs to modify vulnerable words in the input text through synonym replacements. These perturbations have effectively exploited weaknesses in victim models, leading to altered predictions while preserving the original semantic meaning.

The proposed adversarial attack was empirically evaluated on a well-known binary sentiment classification dataset with only single-sentence reviews. In both cases, the word-level perturbation strategy outperformed previous LLM-based state-of-the-art techniques, demonstrating its effectiveness in misleading models.

These results highlight the proposal´s potential as a valuable tool for assessing modern NLP models' robustness and prediction quality. Additionally, it can be integrated into adversarial training frameworks to improve the resilience of NLP systems against adversarial attacks.

Future work should extend the method to include character-level modifications, such as misspellings, letter substitutions, and spacing variations, which can further challenge victim models. Additionally, selecting words for the perturbation task could be refined by incorporating insights from recent literature on word importance. Exploring advanced ranking strategies based on interpretability methods or model attention mechanisms could enhance the effectiveness of the attack while minimizing unnecessary alterations.

ACKNOWLEDGMENT

This research has been supported by grants from the Spanish Ministry of Science and Innovation, under the Knowledge Generation Projects program: XMIDAS (Ref: PID2021-122640OB-100), and the Public-Private Collaboration program: DICYME (Ref: CPP2021-009025).

REFERENCES

[1] I. D. Mienye and T. G. Swart, "A comprehensive review of deep learning: Architectures, recent advances, and applications," *Information*, vol. 15, no. 12, p. 755, 2024.

[2] V. Hassija, V. Chamola, A. Mahapatra, A. Singal, D. Goel, K. Huang, S. Scardapane, I. Spinelli, M. Mahmud, and A. Hussain, "Interpreting black-box models: a review on explainable artificial intelligence," *Cognitive Computation*, vol. 16, no. 1, pp. 45–74, 2024.

[3] Z. Li, M. Wu, C. Jin, D. Yu, and H. Yu, "Adversarial self-training for robustness and generalization," *Pattern Recognition Letters*, vol. 185, pp. 117–123, 2024.

[4] C.-J. H. Yao Li, Minhao Cheng and T. C. M. Lee, "A review of adversarial attack and defense for classification methods," *The American Statistician*, vol. 76, no. 4, pp. 329–345, 2022.

[5] F. Bodria, F. Giannotti, R. Guidotti, F. Naretto, D. Pedreschi, and S. Rinzivillo, "Benchmarking and survey of explanation methods for black box models," *Data Mining and Knowledge Discovery*, vol. 37, no. 5, pp. 1719–1778, 2023.

[6] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang *et al.*, "A survey on evaluation of large language models," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 3, pp. 1–45, 2024.

[7] H. Xu, Y. Ma, H.-C. Liu, D. Deb, H. Liu, J.-L. Tang, and A. K. Jain, "Adversarial attacks and defenses in images, graphs and text: A review," *International journal of automation and computing*, vol. 17, pp. 151–178, 2020.

[8] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, "A survey on adversarial attacks and defences," *CAAI Transactions on Intelligence Technology*, vol. 6, no. 1, pp. 25–45, 2021.

[9] S. Qiu, Q. Liu, S. Zhou, and W. Huang, "Adversarial attack and defense technologies in natural language processing: A survey," *Neurocomputing*, vol. 492, pp. 278–307, 2022.

[10] C. Guo, A. Sablayrolles, H. Jégou, and D. Kiela, "Gradient-based adversarial attacks against text transformers," in *2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 5747–5757.

[11] X. Han, Y. Zhang, W. Wang, and B. Wang, "Text adversarial attacks and defenses: Issues, taxonomy, and perspectives," *Security and Communication Networks*, vol. 2022, no. 1, p. 6458488, 2022.

[12] Z. Wang, W. Wang, Q. Chen, Q. Wang, and A. Nguyen, "Generating valid and natural adversarial examples with large language models," in *27th International Conference on Computer Supported Cooperative Work in Design*, 2024, pp. 1716–1721.

[13] X. Hu, G. Liu, B. Zheng, L. Zhao, Q. Wang, Y. Zhang, and M. Du, "Fasttextdodger: Decision-based adversarial attack against black-box nlp models with extremely high efficiency," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 2398–2411, 2024.

[14] X. Xu, K. Kong, N. Liu, L. Cui, D. Wang, J. Zhang, and M. Kankanhalli, "An LLM can fool itself: A prompt-based adversarial attack," in *12th International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=VVgGbB9TNV

[15] Gemma Team, M. Riviere, S. Pathak, P. G. Sessa, C. Hardin, S. Bhupatiraju, L. Hussenot, T. Mesnard, B. Shahriari, A. Ramé *et al.*, "Gemma 2: Improving open language models at a practical size," 2024.

[16] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, "The llama 3 herd of models," 2024.

[17] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei *et al.*, "Qwen2. 5 technical report," 2024.

[18] A. Young, B. Chen, C. Li, C. Huang, G. Zhang, G. Zhang, G. Wang, H. Li, J. Zhu, J. Chen *et al.*, "Yi: Open foundation models by 01. ai," 2024.

Smart Water Security with AI and Blockchain-Enhanced Digital Twins

1st Mohammadhossein Homaei
Media Engineering Group
University of Extremadura
Cáceres, Spainn
mhomaein@alumnos.unex.es

2nd Víctor González Morales
3rd Óscar Mogollón Gutiérrez
Media Engineering Group
University of Extremadura
Cáceres, Spain
{victorgomo, oscarmg}@unex.es

4th Rubén Molano Gómez
5th Andrés Caro
Media Engineering Group
University of Extremadura
Cáceres, Spain
{rmolano, andresc}@unex.es

***Abstract*—Water distribution systems in rural areas face serious challenges such as a lack of real-time monitoring, vulnerability to cyberattacks, and unreliable data handling. This paper presents an integrated framework that combines LoRaWAN-based data acquisition, a machine learning-driven Intrusion Detection System (IDS), and a blockchain-enabled Digital Twin (BC-DT) platform for secure and transparent water management. The IDS filters anomalous or spoofed data using a Long Short-Term Memory (LSTM) Autoencoder and Isolation Forest before validated data is logged via smart contracts on a private Ethereum blockchain using Proof of Authority (PoA) consensus. The verified data feeds into a real-time DT model supporting leak detection, consumption forecasting, and predictive maintenance. Experimental results demonstrate that the system achieves over 80 transactions per second (TPS) with under 2 seconds of latency while remaining cost-effective and scalable for up to 1,000 smart meters. This work demonstrates a practical and secure architecture for decentralized water infrastructure in under-connected rural environments.**

***Index Terms*—Digital Twins, Blockchain, Cybersecurity, Artificial Intelligence, Intrusion Detection System, Water Industry**

I. Introduction

While remote-sensing techniques have proven valuable for water quality monitoring [1], [2], regarding water distribution, efficient distribution is a significant issue in rural regions, especially where infrastructure is poor and digital monitoring is scarce. Many rural parts of Spain still rely on outdated water distribution systems with manual inspections or partially automated tools, leading to delays in problem detection, inaccurate usage data, and increased risks of human error or data manipulation. Implementing a digital twin (DT) can address these challenges by creating a virtual replica of the water network, enabling operators to detect leaks, predict demand, and optimize maintenance scheduling. However, DT systems also face cybersecurity risks, as data may be altered or falsified before reaching the digital model [3], [4].

DT technology is increasingly central to cyber-physical systems. It provides a real-time virtual representation of physical infrastructures, supporting advanced analysis, forecasting, and anomaly detection [5]–[7]. DTs assist utility operators in identifying leaks, pressure irregularities, and unusual consumption patterns. Despite these advantages, ensuring data security and trustworthiness within DT-based systems remains a significant challenge, particularly in decentralized settings involving multiple stakeholders.

BC technology offers a novel solution for secure data management within online platforms. It ensures decentralized and tamper-proof information storage and verification through Distributed Ledger Technology (DLT), cryptographic security, and consensus mechanisms [4], [8]. Additionally, smart contracts automate tasks like device registration, real-time billing, and fault detection, thus minimizing reliance on intermediaries and increasing accountability [9], [10].

To address these security concerns, our platform incorporates an Artificial Intelligence (AI)-based IDS that identifies anomalous or suspicious data. Only validated and trusted data are forwarded to the BC, where they are securely and permanently stored. Our method integrates LSTM-based anomaly detection and BC technology, ensuring precise, secure, and transparent water management. Without secure, intelligent prediction, water use remains inefficient and decisions are delayed.

This paper presents an integrated DT system supported by BC technology to enhance water distribution management in rural villages in Spain. Data is collected via long-range, low-power LoRaWAN sensors and securely stored on a private BC network, creating an encrypted foundation for the DT. AI and ML techniques enable predictive analytics and anomaly detection, supporting better decision-making and resource optimization. The integration of BC and DT technologies enhances transparency, scalability, and reliability. This cost-effective solution is suitable for rural communities, water utilities, and governmental entities.

The remainder of this paper is structured as follows: Section II reviews related studies on DT and BC applications in water management. Section III describes the proposed framework, including DT architecture, LoRa-based sensor networks, and the integration of a private BC. Section IV evaluates performance, scalability, and security through real-world tests. Section V discusses key findings, challenges, and implications, and concludes by highlighting the contributions and suggesting future research directions aimed at optimizing BC-based DT solutions.

II. Related Works

A. DT in the Water Industry

Recent studies have highlighted the role of DTs as effective tools for enhancing water distribution systems. DT technologies simulate real-time water network behavior, enabling operators to perform leak detection, forecast water consumption, and improve overall maintenance scheduling [5], [11]. For instance, DT platforms have successfully reduced water losses through predictive analytics, proving valuable for improving efficiency and reducing operational costs [6]. However, existing DT implementations often neglect critical cybersecurity considerations, treating data from IoT sensors as inherently trustworthy, which can expose systems to data spoofing and manipulation risks [3].

B. Security in Water Distribution Systems

The digitalization of water systems introduces increased cybersecurity threats, particularly where IoT devices and wireless sensor networks (e.g., LoRaWAN) are extensively deployed. Kim et al. [12] revealed multiple vulnerabilities, including weak encryption methods, poor authentication practices, and outdated communication protocols, leaving these systems susceptible to unauthorized access, data falsification, and denial-of-service (DoS) attacks. Traditional cybersecurity frameworks for water infrastructure tend to reactively detect breaches only after data has been compromised or corrupted, lacking real-time predictive detection capabilities [13]. Consequently, an urgent need exists for proactive anomaly detection systems integrated seamlessly within digital water infrastructures to protect against threats before data reaches critical operational layers like DTs.

C. BC Integration in DT

BC technology has been proposed as a viable solution to improve security, transparency, and immutability in DT applications across multiple sectors. For instance, Mohammed et al. [14] employed Hyperledger Fabric to secure IoT sensor data in smart water management systems, demonstrating enhanced trust and traceability. Similarly, MQTT-based BC solutions have successfully provided tamper-proof logging of sensor readings, reducing risks of data loss and falsification [15]. Despite these advances, existing BC-integrated DT frameworks assume incoming data is valid without verifying it, creating vulnerabilities. Teisserenc et al. [16] addressed some limitations by introducing decentralized DT models with smart contracts for automated decision-making, but lacked robust pre-validation mechanisms to ensure data authenticity. Thus, incorporating anomaly detection before BC data storage is essential to ensure data integrity.

D. AI and ML in IDS for Critical Infrastructure

To overcome limitations of traditional security approaches, AI and ML techniques have become widely adopted for anomaly detection in critical infrastructure, such as energy grids and water systems. Isolation Forest algorithms have successfully identified statistical anomalies in infrastructure

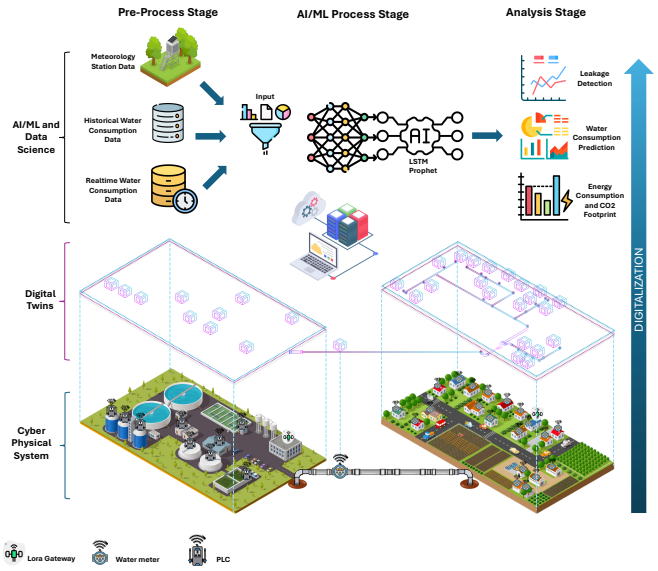


Fig. 1. A Digital Twin Platform in the Water Industry [11]

sensor data, enabling early detection of attacks and faults [8]. Additionally, LSTM Autoencoders have proven effective in detecting temporal anomalies in sequential data, such as abnormal water consumption patterns indicative of leaks or cyberattacks [9], [17]. Nevertheless, existing ML-based IDS solutions are often developed in isolation, lacking integrated deployment within BC-enabled DT frameworks. Furthermore, their evaluation typically occurs under idealized laboratory conditions, without sufficient consideration of intermittent connectivity or rural operational constraints. This research gap motivates the integration of AI-driven IDS within BC and DT ecosystems, specifically addressing rural deployments.

III. Proposed Platform

In this paper, we build on our previous work in the water industry by enhancing the DT system with a stronger focus on security, reliability, and data protection. The DT model proposed in [11] is composed of three main layers (Figure 1): the cyber-physical system (CPS), its digital representation, and an AI-based data analysis and prediction layer. In this updated version, we improve the leak detection process, simplify the identification of data anomalies, and leverage historical data to detect potential cyberattacks and abnormal patterns. To enhance transparency, we integrate a BC system that connects LoRaWAN sensor data to a private Ethereum-based BC. The improved system consists of three key components:

- Leak and unreal detection layer, which processes incoming data and compares it against seasonal trends and long-term average consumption patterns.
- Anomaly and attack detection layer, based on an LSTM model, which prevents out-of-range or incorrect data from being inserted into the BC.

- BC layer, which securely stores validated data from the previous layers on the BC using a dedicated smart contract.

A. Leakage detection

The first step in ensuring water system integrity involves detecting leaks and unrealistic consumption values based on temporal patterns. The platform uses rule-based logic and thresholding derived from long-term consumption profiles to identify potential leaks, especially during non-usage hours (e.g., 00:00–06:00). The logic assumes that consistent water flow during expected inactivity periods likely indicates leakage. Algorithm 1 summarizes this detection process, which serves as a lightweight filtering mechanism before invoking the AI-based IDS for further validation. Leakage alerts are issued locally and passed to the IDS for anomaly confirmation before BC logging.

Algorithm 1 Leakage Detection and Blockchain Validation

Input: LoRaWAN meter data stream D
Output: Leakage alerts, validated BC records
Initialize buffer H, counters
for all $d \in D$ **do**
  Extract hourly data and update H
  Check nighttime (00:00–06:00) consumption
  Update leakage counter: increment if all > 0, else reset
  if counter ≥ 2 **then**
    Flag leakage; freeze status
    if next message confirms leakage **then**
      Alert consumer
    end if
  end if
  Run IDS on d
  if anomaly detected **then**
    Log and reject
  else
    Store on BC
  end if
end for

B. AI-Based IDS for DT

While the BC component guarantees secure and immutable data storage, it does not provide real-time protection against data spoofing, replay attacks, or transaction flooding. To address this, we propose an AI-driven IDS that operates between the data acquisition and BC layers. This IDS employs two complementary models: an LSTM Autoencoder for sequence-based anomaly detection and an Isolation Forest (IF) for statistical outlier detection. Together, they filter out both temporal and point-wise anomalies in water meter data received via LoRaWAN before storing it on the BC.

1) Design and Threat Model: The IDS is designed to detect and block:

- *Spoofed data:* Manipulated readings with plausible structure but inconsistent behavior.
- *Replay attacks:* Repetition of legitimate data to flood or mislead the system.
- *Outliers:* Abnormally high consumption, unexpected error codes, or gas usage irregularities.

While smart contracts verify sender identity and enforce structural rules, they cannot detect logical inconsistencies. The IDS addresses this gap through both statistical and temporal pattern learning.

Algorithm 2 Combined LSTM and Isolation Forest IDS

Input: Trained LSTM model $\mathcal{M}$, trained IF model $\mathcal{F}$, thresholds τ, θ
For each meter: maintain buffer $\mathbf{X}_m$ of size N
for all incoming event e_t **do**
  Extract feature vector $\mathbf{x}_t$ and append to $\mathbf{X}_m$
  Compute IF anomaly score: $s \leftarrow \mathcal{F}(\mathbf{x}_t)$
  if $s > \theta$ **then**
    Raise anomaly alert (Isolation Forest)
    Reject record and log the incident
  else if $|\mathbf{X}_m| = N$ **then**
    $\hat{\mathbf{X}}_m \leftarrow \mathcal{M}.decode(\mathcal{M}.encode(\mathbf{X}_m))$
    Compute reconstruction loss $\mathcal{L}_{\text{recon}}$
    if $\mathcal{L}_{\text{recon}} > \tau$ **then**
      Raise anomaly alert (LSTM Autoencoder)
      Reject record and log incident
    else
      Accept and forward to BC
    end if
    Remove oldest vector from $\mathbf{X}_m$
  end if
end for

2) Feature Engineering: The IDS continuously processes incoming real-time data streams from LoRaWAN sensors and BC logs. For each new record, a feature vector is constructed as:

$$\mathbf{x}_t = [\text{WaterUsage}_t, \text{ErrorCode}_t, \text{TxRate}_t, \text{GasUsed}_t] \quad (1)$$

For the LSTM model, a sequence of N such vectors is maintained per meter:

$$\mathbf{X}_m = \{\mathbf{x}_{t-N+1}, \ldots, \mathbf{x}_t\} \quad (2)$$

3) LSTM Autoencoder Architecture: The LSTM autoencoder learns to reconstruct sequences of normal behavior. It encodes a sequence into a latent representation and reconstructs it, allowing anomaly detection via reconstruction error:

$$\mathcal{L}_{\text{recon}} = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|^2 \quad (3)$$

If $\mathcal{L}_{\text{recon}} > \tau$, where τ is a predefined threshold, the sequence is flagged as anomalous.

4) Isolation Forest Outlier Detection: To complement the LSTM, we use an Isolation Forest trained on individual features to detect non-sequential outliers [18]. Given a new observation $\mathbf{x}_t$, the IF model returns an anomaly score $s(\mathbf{x}_t)$ based on how easily the point is isolated in the tree ensemble. An alert is raised if:

$$s(\mathbf{x}_t) > \theta \quad (4)$$

where θ is the anomaly threshold determined during training.

5) Real-Time Detection Algorithm: The detection process operates in real time using a sliding window buffer and event-based triggers from smart contracts. A record must pass both LSTM-based sequence analysis and IF outlier detection before being accepted.

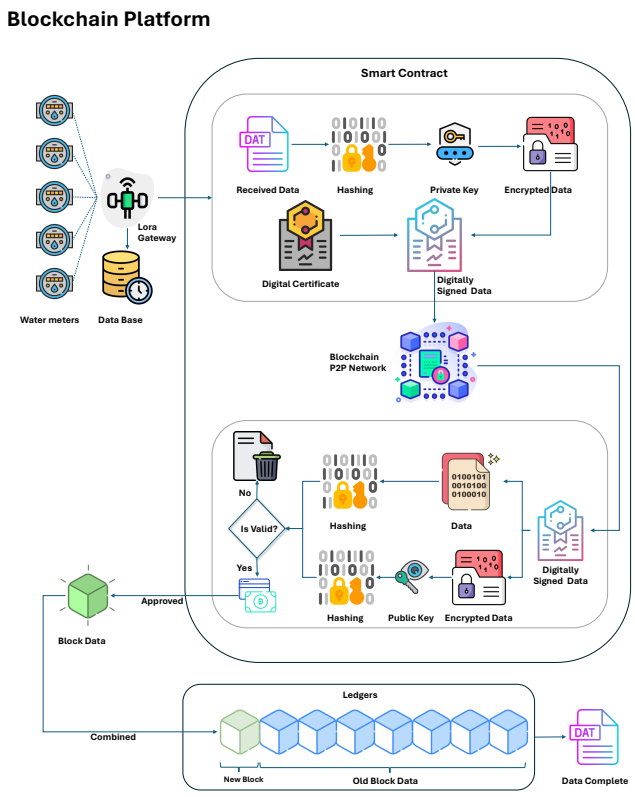


Fig. 2. Proposed Smart contract for DT platform

6) Integration with Blockchain: The IDS listens to smart contract events (e.g., *WaterDataLogged*) via Web3 interfaces and buffers incoming records accordingly. Its output dictates whether the data is stored on the BC or discarded. Optional logging of detected anomalies on-chain can improve transparency, support audits, and train future models. The IDS layer is modular and can be deployed alongside any BC network or smart contract design, ensuring seamless integration and compatibility with decentralized infrastructures.

C. BC

Figure 2 illustrates the smart contract structure used in the proposed platform. This contract handles essential functions such as meter registration, secure logging of consumption data, and automatic transaction processing. Specifically, the contract defines a 'WaterData' structure that records the timestamp, water usage, error codes, and associated meter IDs. Functions such as *registerMeter()*, *logWaterData()*, and *calculatePayment()* are designed to enforce access control, validate data, and automate billing based on consumption and error codes, as detailed in Algorithm 3.

In parallel, Figure 3 presents the core technologies involved in the BC layer. This includes the use of a private Ethereum network with a PoA consensus model, which ensures fast finality and minimal energy consumption, particularly suitable for edge computing scenarios in rural environments. The integration of DTs with smart contracts enables not only secure

Algorithm 3 DT and BC Smart Contract

```
Contract VillageWaterSystem
Struct WaterData: uint256 timestamp, waterUsage, errorCode; string meterId
address owner; mapping(string =¿ WaterData[]) waterLogs; string[] registeredMeters
Event MeterRegistered, MeterDisabled, WaterDataLogged, PaymentProcessed
function ONLYOWNER
    Require(msg.sender = owner, "Unauthorized")
end function
function CONSTRUCTOR VILLAGEWATERSYSTEM
    owner ← msg.sender
end function
function REGISTERMETER(string meterId)
    Require(meterId ≠ empty, "Invalid ID")
    registeredMeters.push(meterId); Emit MeterRegistered(meterId)
end function
function DISABLEMETER(string meterId)
    Require(meterId ≠ empty, "Invalid ID")
    Remove from registeredMeters; Emit MeterDisabled(meterId)
end function
function LOGWATERDATA(string id, uint256 u, uint256 e)
    Require(isMeterRegistered(id), "Unreg.")
    Require(e ≤ 100, "Invalid err")
    Store in waterLogs[id]; Emit WaterDataLogged(id, u, e)
    p ← calculatePayment(u, e); Emit PaymentProcessed(id, p)
end function
function ISMETERREGISTERED(string id) returns bool
    Return (id in registeredMeters)
end function
function CALCULATEPAYMENT(uint256 u, e) returns uint256
    Return u * 1 ether * ((e > 80) ? 1 : 2)
end function
function GETWATERLOGS(string id) returns WaterData[]
    Require(isMeterRegistered(id), "Unreg.")
    Return waterLogs[id]
end function
function GETREGISTEREDMETERS returns string[]
    Return registeredMeters
end function
function SETBASERATE(uint256 r) onlyOwner
    Require(r > 0, "Invalid rate")
end function
```

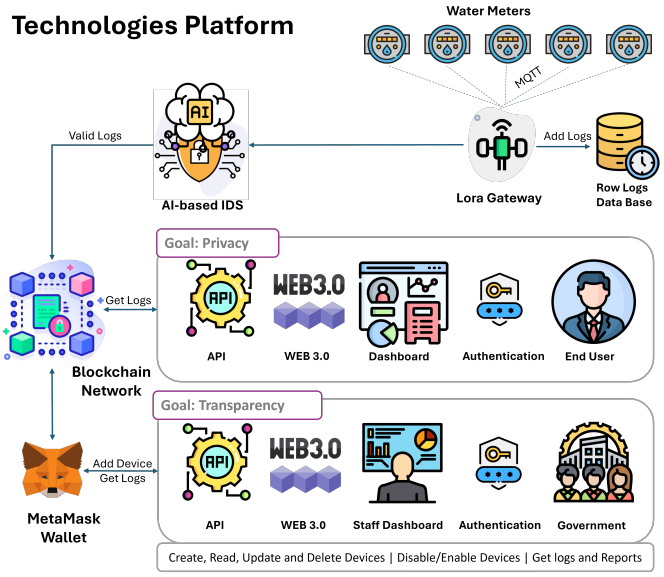


Fig. 3. Technologies in the Platform on the BC side

data storage but also autonomous system behavior, reducing reliance on central servers or human intervention. By combining BC, DT, and AI-driven verification mechanisms, the platform offers a scalable, transparent, and resilient solution for water resource management in under-connected areas.

IV. Evaluation

This section presents a practical evaluation of the proposed framework under conditions typical of rural Spanish villages. We assess system-level performance—including throughput, latency, and scalability—as well as security and deployment cost. A private Ethereum blockchain with PoA consensus was

deployed on a dedicated Hetzner server, and a LoRaWAN metering environment was used to emulate real-time consumption data. The evaluation includes analysis of the hardware/software setup, network topology, anomaly detection, tamper resistance, and cost-efficiency.

A. Leakage Detection Results

To verify the effectiveness of the leakage detection mechanism, we analyzed historical water meter data from 400 devices deployed across rural locations for last three years. The detection algorithm flagged meters with non-zero night consumption over consecutive days, suggesting probable leaks.

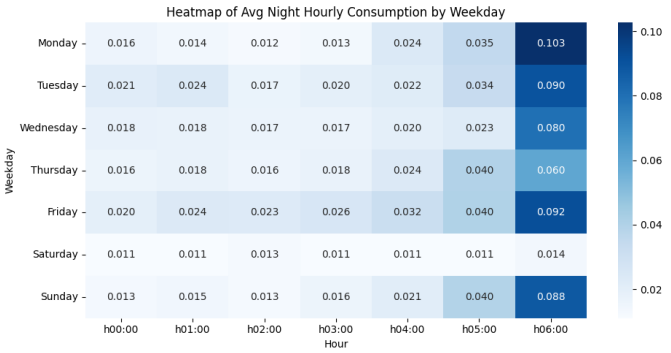


Fig. 4. Night consumption Hitmap for a water meter with leakage

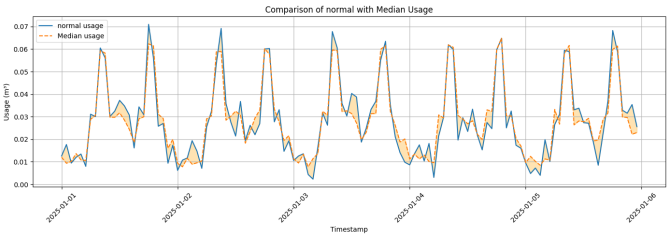


Fig. 5. Comparison Normal and Median usage

Figure 4 shows a heatmap of night-time water consumption (00:00–06:00) for a leaking meter, where consistent activity was detected. Figure 5 compares normal consumption patterns to the median usage, highlighting that anomalies often deviate from expected seasonal or diurnal trends. Additionally, Figure 6 aggregates the night usage across all flagged meters, reinforcing the accuracy of the detection logic.

B. Anomaly Detection Result via IDS

The system was evaluated using real-world water consumption data injected with synthetic attacks. Figure 7 shows a direct comparison between a normal consumption pattern and an anomalous one. The sample line represents typical usage behavior, while the pattern line reveals injected anomalies, such as abnormal spikes and repeated low-consumption values that mimic night-time leakage or spoofed records.

To further emphasize deviations, Figure 8 plots the detected anomalous consumption values against the meter's typical

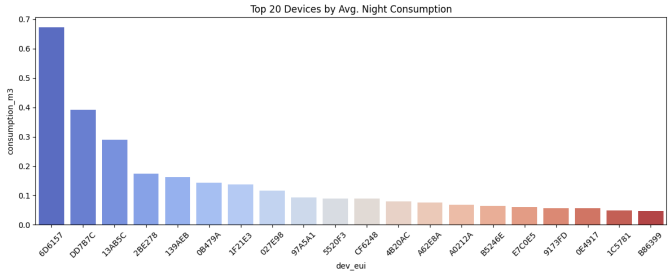


Fig. 6. Leaked meters and their consumption during nights

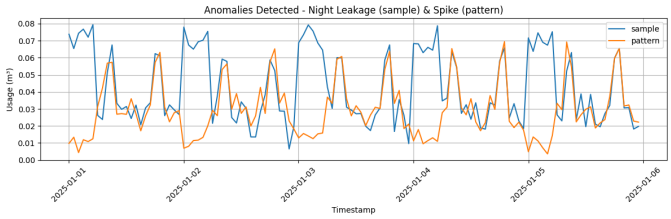


Fig. 7. Anomalies Detected: Comparison of Night Leakage (sample) and Spike Patterns

median usage. The sharp divergence observed during the attack period clearly demonstrates how the IDS identifies data points that deviate significantly from normal patterns while maintaining temporal consistency.

Figure 9 presents a heatmap comparing the IDS performance metrics—Precision, Recall, and F1-Score—for each type of attack. This visual summary is consistent with the quantitative results shown in Table I, highlighting the IDS's effectiveness across diverse anomaly types.

TABLE I
ANOMALY DETECTION RESULTS OF THE HYBRID IDS

Attack Type	Injected	Detected	Precision	Recall	F1-Score
Replay Attack	120	112	0.93	0.93	0.93
Spoofed Consumption	100	97	0.91	0.97	0.94
Tampered Error Codes	80	76	0.89	0.95	0.92
Gas Usage Anomalies	70	64	0.91	0.91	0.91
Overall	370	349	**0.91**	**0.94**	**0.92**

C. Experimental Setup BC

- **Hardware and Software Configuration:** The proposed framework was deployed on a Hetzner server running a private Ethereum network with a PoA

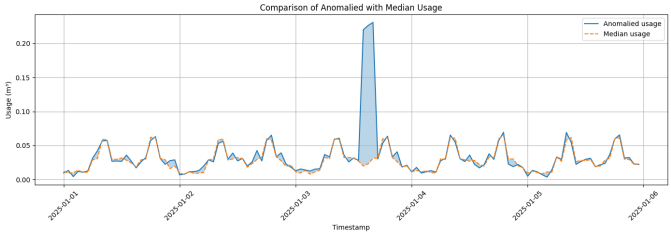


Fig. 8. Detected Anomalies Compared with Median Usage Baseline

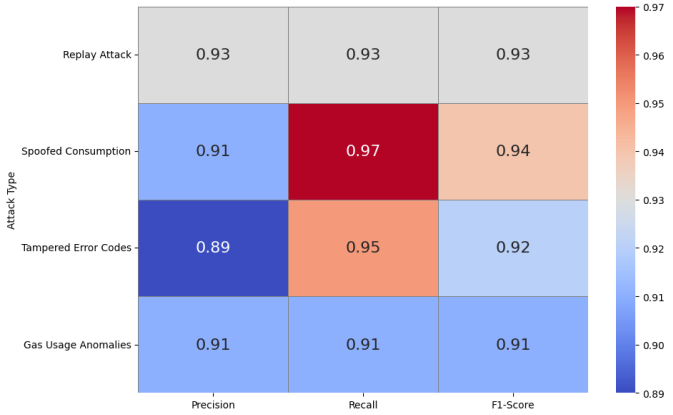


Fig. 9. Heatmap of Precision, Recall, and F1-Score per Attack Type

consensus mechanism. Three validator nodes operated via Docker containers to simulate on-chain validation. Meter data from 400 emulated LoRaWAN devices was batched and submitted in 8-hour intervals. Prometheus and Grafana were used for monitoring. The setup reflects real-world rural conditions, tolerating intermittent connectivity while ensuring secure, scalable, and low-latency operation.

- **Network Topology:**
  - *Validator Nodes:* Each validator node runs with a 1-second block interval and a block gas limit of 15 million—settings that enable higher throughput and faster transaction finalization than default Ethereum configurations. We deployed three validator nodes on the same Hetzner dedicated server, each operating in its own Docker container. The nodes communicate over a secure internal network and use static peer discovery to maintain connectivity and validate on-chain transactions.
  - *LoRaWAN Gateway:* Meter data is transferred to the BC through a replicated LoRaWAN gateway that aggregates sensor data every 8 hours. It then batches these readings into transactions before submitting them on-chain, ensuring any intermittent connectivity does not result in data loss.
- **Water Metering Scenario:** In our test scenario, we emulate the behavior of 400 water meters installed across multiple rural areas. Each meter is configured to capture the following data at 8-hour intervals:
  - Meter ID
  - Timestamp
  - Water Consumption (in cubic metres)
  - Error Code

  Each meter transmits its data three times per day to account for potential downtime or connectivity issues. This setup reflects the actual operational conditions in remote Spanish villages, where consistent Internet access cannot always be guaranteed.

D. Performance and Scalability

In this section, we present the results of our performance analysis and scalability tests for the proposed BC-based DT framework. We aim to demonstrate that the system can process meter readings efficiently, maintain low transaction latency, and scale to accommodate an increasing number of water meters.

- **Transaction Throughput and Latency:**
  To evaluate the performance of the BC network, we focused on two core metrics:
  - *TPS's Throughput:* The number of successfully confirmed transactions per second.
  - *Transaction Latency (Seconds):* The time interval between the client submitting the transaction and its final confirmation on-chain.

  We conducted a series of stress tests by varying the batch sizes (i.e., how many meter readings are grouped into a single on-chain transaction). This approach allowed us to evaluate the system's behavior under different data aggregation strategies—especially relevant for rural deployments where intermittent connectivity may lead to buffered uploads.
  The Mean Latency in Table II refers to the time from when a transaction is submitted to its first inclusion in a block. Finality in our PoA setup typically arrives 1–2 blocks after inclusion, corresponding to an additional 2–3 seconds under typical loads.

TABLE II
TRANSACTION THROUGHPUT AND LATENCY UNDER DIFFERENT BATCHING CONDITIONS

Batch Size	Meters Tested	Throughput (TPS)	Mean Latency (s)	Max Latency (s)
1 reading/tx	400	110	1.2	2.1
5 readings/tx	400	96	1.5	2.4
10 readings/tx	400	89	1.7	2.8
20 readings/tx	400	81	2.1	3.5

  Observations:
  - As the batch size increases, throughput decreases slightly, attributable to larger transaction payloads requiring more on-chain processing time.
  - Latency grows proportionally with the batch size. However, even at 20 readings per transaction, the network sustains an average throughput of 81 TPS with a mean latency of around 2 seconds.
  - These results indicate that our PoA network can efficiently handle data bursts from hundreds of meters, making it suitable for real-world deployments where large numbers of sensors may periodically transmit readings.
- **Block Finalization:**
  Using a PoA consensus mechanism provides faster block finalization times compared to traditional Proof of Work (PoW) networks. Our experiments show that blocks are typically finalized within **2–3 seconds** under the tested workloads. This quick finality has two primary benefits:

1) *Timely Data Recording:* Water consumption data are confirmed on-chain almost immediately, enabling near real-time monitoring within the DT environment.
2) *Predictive Maintenance:* Rapid confirmation aids anomaly detection algorithms in swiftly identifying irregularities (e.g., leaks or sensor malfunctions), reducing response times and potential water losses.

Such short block finalization intervals are particularly valuable for rural water management, where operators rely on accurate, up-to-date information to schedule maintenance tasks, plan usage patterns, and optimize resources.

- **Scalability with Increasing Meter Count:**
  To assess how the system behaves under a growing number of sensors, we conducted additional tests by gradually scaling the number of simulated meters from 100 to 1,000 while holding the transaction batch size at five readings per transaction. Across these tests:
  - The network maintained a throughput greater than 85 TPS in all experiments, even as the meter count increased by an order of magnitude.
  - System latency showed minimal growth, reinforcing the notion that the PoA-based framework scales well with increased demand.
  - These results underscore the system's potential to extend to larger water distribution networks without significant performance degradation, making it suitable for both small rural communities and larger municipal deployments.

E. Security and Reliability

Security and reliability are central pillars of any data management solution for critical infrastructure like water distribution. BC's immutable ledger and PoA-based access controls work together to ensure that the system is tamper-resistant and fault-tolerant. Below, we detail the measures taken to protect against unauthorized modifications and to maintain network reliability.

- **Data Immutability and Tamper Resistance:**
  One of the chief advantages of a BC-based solution is the *immutability* of on-chain records. We conducted targeted tests to confirm that malicious attempts to alter data or inject bogus information would be rejected:
  - *Direct Database Manipulation:* We tried modifying the raw on-chain data files stored in the local node's directory. The PoA consensus nodes detected mismatched hashes, invalidating the altered data.
  - *Smart Contract Override:* We attempted to call special administrative functions like *logWaterData* and *disableMeter* without proper credentials. These calls were blocked by role-based access controls enforced at the smart contract level.
  - *Spurious Node Injection:* We introduced a rogue node with a manipulated ledger history, which the existing validator nodes refused to add to the network.

  Table III summarizes the outcome of these tests:

TABLE III
TAMPER-RESISTANCE TEST RESULTS

Attempted Attack	Result
Direct on-disk data modification	Rejected (immutable ledger)
Unauthorized smart contract invocation	Rejected (access control)
Rogue validator introduction	Blocked (PoA authority management)

  All unauthorized modifications were invalidated by the network's consensus protocol, confirming that meter data remains tamper-proof once recorded on-chain. This reliability is crucial for building trust among municipalities, local water authorities, and end-users.
- **Access Control and Authentication:**
  Access control is enforced via smart contracts. Before a meter can submit data, it must be registered on-chain by an authorized administrator. We tested unauthorized submissions to assess whether the system would correctly reject them:
  - *Fake Meter ID:* A transaction with an unregistered meter ID triggered an immediate rejection in the *isMeterRegistered* function.
  - *Valid Meter ID, Incorrect Credential:* If the transaction was signed by a private key not recognized by the PoA nodes, the network discarded the transaction before it reached the contract logic.

  These findings confirm that the framework effectively prevents unauthorized data entries and ensures only valid meter readings are integrated into the DT environment.
- **Network Reliability in Rural Deployments:**
  Rural Spanish villages often face intermittent Internet connectivity. Our solution, therefore, tolerates temporary offline periods without losing data integrity. We simulated a scenario with a 2-hour daily connectivity loss over a week:
  - The LoRaWAN gateway buffered the readings until the BC node was reachable.
  - Upon reconnection, the pending transactions were submitted in batches.
  - No BC reorganization occurred, as the PoA validators incorporated the newly arrived batches without conflict.

  This experiment demonstrates the system's resilience, confirming that brief outages do not compromise the integrity or completeness of stored data. Such robustness is essential for real-world deployments where continuous high-speed Internet is not always available.

F. Cost Analysis

Although public BCs typically require gas fees for each transaction, our private PoA network could be configured to impose negligible or zero gas costs, significantly reducing financial overhead for municipalities. Our PoA nodes are

hosted on a dedicated Hetzner server, with monthly costs ranging between €20 and €50, depending on the chosen plan. BC maintenance is also cost-effective, as PoA consensus eliminates CPU-intensive mining tasks, and validator nodes have minimal resource requirements beyond basic computation and storage. In terms of network traffic, LoRaWAN-to-BC communications incur only minor data charges, while on-chain transaction fees can be set to near zero, avoiding high per-transaction costs. The system's scalability ensures that adding additional meters does not significantly increase operational expenses since the same validator nodes can efficiently handle larger data volumes within tested limits. Furthermore, the architecture's linear scalability ensures that even large-scale deployments remain affordable. Table IV provides an approximate breakdown of costs for a six-month pilot project, excluding expenses related to physical LoRaWAN gateways or sensors, which vary based on specific deployment needs.

TABLE IV
COST ESTIMATION BREAKDOWN IN A 6-MONTH PILOT

Component	Cost (EUR)	Notes
Server Rental (6 months)	120–300	Depends on hosting plan
Maintenance	˜50	Occasional reboots, software updates
Energy	Included	Covered by hosting service
LoRaWAN Gateways	Variable	Based on deployment size and hardware choice
On-chain Gas Fees	Near-zero	PoA network with custom gas price

Overall, this PoA-based BC solution is cost-effective for municipalities of varying sizes, especially when compared to traditional centralized data management systems that may involve higher maintenance and licensing fees.

V. CONCLUSION

This paper introduced a Blockchain-based Digital Twin (BC-DT) framework that combines a private PoA Ethereum blockchain, LoRaWAN sensors, and a hybrid Intrusion Detection System using LSTM Autoencoder and Isolation Forest to enhance the security and reliability of rural water distribution systems. The proposed system enables real-time anomaly detection, secure data logging via smart contracts, and supports transparent, decentralized monitoring. Evaluation results demonstrated strong performance with over 80 TPS, low latency, tamper resistance, and cost-effective scalability across 1,000 smart meters. The architecture is resilient to intermittent connectivity and adaptable to rural infrastructure constraints. Future work will explore enhancements such as federated learning for decentralized model training, dynamic pricing via smart contracts, and energy-efficient scaling to urban and industrial settings.

ACKNOWLEDGMENT

This initiative is carried out within the framework of the funds from the Recovery, Transformation, and Resilience Plan, financed by the European Union (Next Generation) – National Institute of Cybersecurity (INCIBE), as part of project C107/23: "Artificial Intelligence Applied to Cybersecurity in Critical Water and Sanitation Infrastructures."

REFERENCES

[1] A. Cuartero, J. Cáceres-Merino, and J. A. Torrecilla-Pinero, "An application of c2-net atmospheric corrections for chlorophyll-a estimation in small reservoirs," *Remote Sensing Applications: Society and Environment*, vol. 32, p. 101021, 2023.

[2] J. Cáceres Merino, A. Cuartero Sáez, and J. Á. Torrecilla Pinero, "Finding optimal spatial window: the influence of size on remote-sensing-based chl-a prediction in small reservoirs," *IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING*, vol. 17, p. 18769, 2024.

[3] A. Alshami, E. Ali, M. Elsayed, A. E. E. Eltoukhy, and T. Zayed, "Iot innovations in sustainable water and wastewater management and water quality monitoring: A comprehensive review of advancements, implications, and future directions," *IEEE Access*, vol. 12, p. 58427–58453, 2024.

[4] M. Homaei, O. Mogollón-Gutiérrez, J. C. Sancho, M. Ávila, and A. Caro, "A review of digital twins and their application in cybersecurity based on artificial intelligence," *Artificial Intelligence Review*, vol. 57, no. 8, Jul. 2024. [Online]. Available: http://dx.doi.org/10.1007/s10462-024-10805-3

[5] W. Li, Z. Ma, J. Li, Q. Li, Y. Li, and J. Yang, "Digital twin smart water conservancy: Status, challenges, and prospects," *Water*, vol. 16, no. 14, p. 2038, Jul. 2024.

[6] S. R. Krishnan, M. K. Nallakaruppan, R. Chengoden, S. Koppu, M. Iyapparaja, J. Sadhasivam, and S. Sethuraman, "Smart water resource management using artificial intelligence—a review," *Sustainability*, vol. 14, no. 20, p. 13384, Oct. 2022.

[7] M. H. Homaei, A. C. Lindo, J. C. S. Núñez, O. M. Gutiérrez, and J. A. Díaz, "The role of artificial intelligence in digital twin's cybersecurity," in *Proceedings of the RECSI - Reunión Española sobre Criptología y Seguridad de la Información*, vol. 6. Spain: RECSI, 2022, p. 7.

[8] A. Fuller, Z. Fan, C. Day, and C. Barlow, "Digital twin: Enabling technologies, challenges and open research," *IEEE Access*, vol. 8, pp. 108 952–108 971, 2020.

[9] D. Kirli, B. Couraud, V. Robu, M. Salgado-Bravo, S. Norbu, M. Andoni, I. Antonopoulos, M. Negrete-Pincetic, D. Flynn, and A. Kiprakis, "Smart contracts in energy systems: A systematic review of fundamental approaches and implementations," *Renewable and Sustainable Energy Reviews*, vol. 158, p. 112013, Apr. 2022.

[10] T. K. Satilmisoglu, Y. Sermet, M. Kurt, and I. Demir, "Blockchain opportunities for water resources management: A comprehensive review," *Sustainability*, vol. 16, no. 6, p. 2403, Mar. 2024.

[11] M. Homaei, A. J. Di Bartolo, M. Ávila, O. Mogollón-Gutiérrez, and A. Caro, "Digital transformation in the water distribution system based on the digital twins concept," 2024. [Online]. Available: https://arxiv.org/abs/2412.06694

[12] E. Kim, "Ensuring cybersecurity in water distribution networks: a risk-based approach," *Journal of Water Resources Planning and Management*, vol. 147, no. 9, 2021.

[13] D. Park and H. You, "A digital twin dam and watershed management platform," *Water*, vol. 15, no. 11, p. 2106, Jun. 2023.

[14] M. A. Mohammed, A. Lakhan, K. H. Abdulkareem, M. K. Abd Ghani, H. A. Marhoon, S. Kadry, J. Nedoma, R. Martinek, and B. G. Zapirain, "Industrial internet of water things architecture for data standarization based on blockchain and digital twin technology," *Journal of Advanced Research*, vol. 66, p. 1–14, Dec. 2024.

[15] M. Naqash, T. Syed, S. Alqahtani, M. Siddiqui, A. Alzahrani, and M. Nauman, "A blockchain based framework for efficient water management and leakage detection in urban areas," *Urban Science*, vol. 7, no. 4, p. 99, Sep. 2023.

[16] B. Teisserenc and S. Sepasgozar, "Adoption of blockchain technology through digital twins in the construction industry 4.0: A pestels approach," *Buildings*, vol. 11, no. 12, p. 670, Dec. 2021.

[17] O. M. Gutiérrez, J. C. S. Núñez, M. Homaei, and J. A. Díaz, "Aplicación de técnicas de reducción de dimensionalidad y balanceo en ciberseguridad," in *VII Jornadas Nacionales de Investigación en Ciberseguridad (JNIC)*, Bilbao, Spain, June 2022.

[18] B. E. Downey, C. K. Leung, A. G. M. Pazdor, R. A. L. Petrillo, D. Popov, and B. R. Schneider, *Anomaly Detection with Generalized Isolation Forest*. Springer Nature Switzerland, 2024, p. 356–368.

DemoTwins Project: Promoting Digital Twins through a Demonstration Center in Extremadura

1st Daniel Flores-Martin
COMPUTAEX.
Extremadura Supercomputing Center
Cáceres, Spain
daniel.flores@computaex.es

2nd Felipe Lemus-Prieto
COMPUTAEX.
Extremadura Supercomputing Center
Cáceres, Spain
felipe.lemus@computaex.es

3rd Noelia Alonso
COMPUTAEX.
Extremadura Supercomputing Center
Cáceres, Spain
noelia.alonso@computaex.es

4th Anthony Farroñan
COMPUTAEX.
Extremadura Supercomputing Center
Cáceres, Spain
anthony.farronan@computaex.es

5th Juan A Rico-Gallego
COMPUTAEX.
Extremadura Supercomputing Center
Cáceres, Spain
juanantonio.rico@computaex.es

***Abstract*—Digital Twin technology has emerged as a transformative innovation, combining artificial intelligence, data analytics, real-time sensors, and advanced simulation techniques to create highly accurate virtual replicas of physical entities and systems. These virtual models enable continuous monitoring, predictive analysis, and scenario simulation, significantly enhancing decision-making processes across various sectors. Despite its vast potential, societal awareness and adoption remain limited, especially among smaller communities and enterprises. This work presents the DemoTwins project, a dedicated Demonstration Center developed by COMPUTAEX in Extremadura, Spain. This center serves as an accessible platform for stakeholders—including businesses, educational institutions, and public organizations—to gain hands-on experience, technical assistance, and practical knowledge about digital twins.**

***Index Terms*—Digital twins, Technologies, Use-case, Demonstration Center**

I. Introduction

Artificial Intelligence (AI) and associated emerging technologies have significantly influenced industry transformations by improving decision-making processes, enabling automation, and enhancing predictive capabilities. A key technological advancement within this domain is the concept of digital twins, which integrates real-time data, analytical methods, and simulation techniques to create highly precise digital replicas of physical objects, systems, or processes [6].

Digital twins provide substantial benefits such as improved operational efficiency, the ability to conduct predictive maintenance, cost optimization, and better resource management [8]. By enabling stakeholders to predict and analyze potential scenarios, digital twins help mitigate risks and optimize processes without disrupting real-world operations. These capabilities are applicable across various fields, including industrial manufacturing, urban planning, agriculture, infrastructure management, healthcare services, automotive production, and environmental sustainability initiatives. Additionally, digital twins facilitate detailed analyses of complex systems, effectively bridging the gap between theoretical models and practical implementations.

However, despite their considerable potential, digital twins are not yet widely recognized or adopted, particularly among small and medium-sized enterprises (SMEs) and local communities. It is therefore essential to increase general awareness and understanding of this technology through practical demonstrations [1].

This work presents the DemoTwins project [1], a project that addresses this critical issue by establishing the Extremadura Digital Twin Demonstration Center, designed specifically to illustrate real-world digital twin applications. The center aims to educate stakeholders and accelerate the broader adoption of digital twin technology by providing concrete examples, direct engagement, and expert guidance, thereby fostering technological innovation and industry advancement.

The rest of the paper is structured as follows: Section II presents the background and motivations for this project. Section III details the project's main objective and key use cases. Finally, Section IV summarizes the conclusions and future work.

II. Background and Motivations

Despite the numerous advantages of digital twin technology, its awareness and adoption remain limited, especially in smaller communities and among SMEs. There is an urgent need to bridge this knowledge gap by demonstrating tangible benefits and facilitating hands-on interaction with the technology [5].

Several international projects and initiatives have aimed at promoting digital twin technologies, such as the Digital Twin Consortium [3], the National Digital Twin Programme in the UK [7], and the Smart Cities initiatives across Europe [2]. In Extremadura, despite growing interest from various sectors—such as agriculture, urban planning, and infrastructure

[1]https://demotwins.computaex.es

management—the adoption of digital twin technology has been hindered by a lack of accessible information and practical examples.

The DemoTwins project addresses this need by establishing a dedicated Demonstration Center as a local resource for education, collaboration, and technological experimentation [4]. This center directly supports regional actors by providing practical demonstrations, expert consultations, and a collaborative environment designed to overcome informational and technological barriers, thus accelerating the adoption of digital twins in Extremadura.

III. Extremadura Digital Twin Demonstrator Center

The central element of the DemoTwins project is the creation of the Extremadura Digital Twin Demonstration Center. This allows functions as an accessible platform for individuals, businesses, and institutions interested in exploring digital twin technology, viewing practical applications, and obtaining specialized technical assistance. In addition to the digital twins developed during the project, interested companies with their digital twin can disseminate them through the demonstration center, either in person or virtually.

Currently, the center highlights interactive digital twin applications across three strategic sectors:

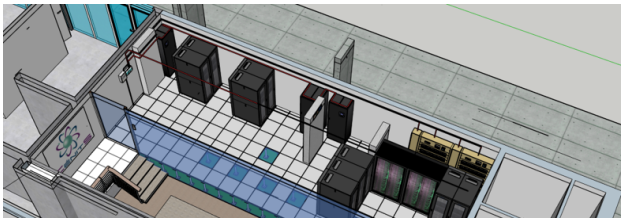
- **Urban management:** digital twins integrated with geographic information systems (GIS) to facilitate urban planning, sustainability initiatives, and smart city developments.
- **Agro-Industrial sector:** digital twins designed to optimize agricultural processes, enhance resource management, and increase productivity through real-time analytics and predictive capabilities.
- **Infrastructure management:** digital twins intended to improve maintenance practices, safety standards, and resilience within critical infrastructure systems such as transportation, energy, and utilities.

The center offers additional digital twin prototypes addressing other sectors, broadening societal awareness and technology adoption. These interactive models employ advanced visualization and modeling technologies, including SketchUp, Unity, Matterport, Revit, and GIS tools. The utilization of these platforms increases the realism, usability, and practical relevance of the demonstrations, enhancing the visitors' experience and understanding. Figure 1 shows an example of the digital twin of the Extremadura Supercomputing Center.

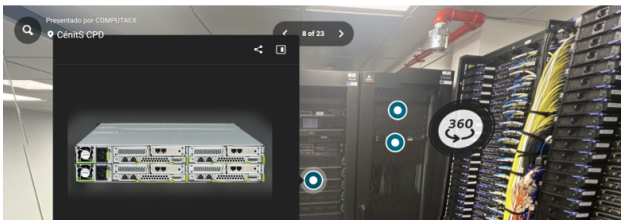
Visitors can benefit from exploring realistic prototypes, obtaining personalized insights into implementation strategies, and receiving expert support, thereby stimulating innovation and technology transfer.

IV. Conclusions

Digital Twin technology represents a significant advancement capable of reshaping industries by optimizing performance and efficiency. However, widespread adoption requires greater awareness and accessibility through practical demonstrations.



(a) BIM Model (Sketchup)



(b) Virtual tour (Matterport)

Fig. 1: COMPUTAEX Processing Center digital twin overview

The DemoTwins project, via the Extremadura Digital Twin Demonstration Center, actively contributes to this awareness and educational effort.

Through direct engagement, practical examples, and expert guidance, DemoTwins enables organizations and communities to recognize and harness the potential advantages of digital twins, laying the foundation for more intelligent, sustainable, and efficient systems and processes in the future.

Acknowledgment

This work was supported by the Ministry of Economy, Employment and Digital Transformation of Junta de Extremadura (Spain) (project 5041, code MR05C13I01). We also thank the COMPUTAEX Foundation for allowing us to use the computational resources of the LUSITANIA supercomputer.

References

[1] Broo, D.G., Schooling, J.: Digital twins in infrastructure: definitions, current practices, challenges and strategies. International Journal of Construction Management **23**(7), 1254–1263 (2023)

[2] Commission, E.: City initiatives. https://commission.europa.eu/eu-regional-and-urban-development/topics/cities-and-urban-development/city-initiatives/smart-cities_en, accessed March 25, 2025

[3] Consortium, D.T.: Digital twin consortium. https://www.digitaltwinconsortium.org, accessed March 25, 2025

[4] de Extremadura, J.: The regional government is betting on the potential of the technology known as 'digital twins' to boost extremadura's industry. https://www.juntaex.es/w/jornada-gemelos-digitales, accessed March 25, 2025

[5] Ferré-Bigorra, J., Casals, M., Gangolells, M.: The adoption of urban digital twins. Cities **131**, 103905 (2022)

[6] Grieves, M.W.: Digital twins: past, present, and future. In: The digital twin, pp. 97–121. Springer (2023)

[7] (NDTP), N.D.T.P.: National digital twin programme (ndtp). https://www.gov.uk/government/collections/the-national-digital-twin-programme-ndtp, accessed March 25, 2025

[8] Sjarov, M., Lechler, T., Fuchs, J., Brossog, M., Selmaier, A., Faltus, F., Donhauser, T., Franke, J.: The digital twin concept in industry–a review and systematization. In: 2020 25th IEEE international conference on emerging technologies and factory automation (ETFA). vol. 1, pp. 1789–1796. IEEE (2020)

Code Deobfuscation Using Chatbots

Joana Moreira
Department of Informatics
Universidade da Beira Interior
Portugal

Vasco Lopes
NOVA LINCS
Universidade da Beira Interior
Portugal

Pedro R. M. Inácio
Instituto de Telecomunicações
Universidade da Beira Interior
Portugal

***Abstract*—The growing accessibility of large language models through public chatbots has opened a large number of new possibilities for code comprehension tasks. In this paper, we study the use of publicly available chatbots for the challenge of code deobfuscation. The core objective is to evaluate the extent to which free, general-purpose large language models (LLMs) can assist in reversing obfuscated code into a human-readable and commented source code. This approach is motivated by the extensive training data that LLMs have access to and the potential gains in productivity for software developers. For this purpose, we propose a testing framework designed to assess whether these chatbots can effectively deobfuscate code, perform reverse engineering, and to what degree of accuracy and consistency this process can be achieved. Moreover, we evaluate whether this process can be performed without requiring advanced domain-specific expertise, with results indicating that well-crafted prompts with contextual examples significantly improve deobfuscation success.**

***Index Terms*—Code Obfuscation, Code Deobfuscation, Chatbots, Large Language Models (LLMs)**

I. Introduction

In recent years, the use of chatbots powered by large language models (LLMs) has become increasingly common in software development [1]. Developers now rely on the interfaces for code generation, explanation, and debugging making them a big part of modern programming [1]. One area that can be further explored is the use of LLMs to reverse code obfuscation [2]. Obfuscation is a technique that alters the structure of the source code to make it harder to understand while deobfuscation attempts to convert obfuscated code into a more readable and logically coherent form. Traditionally, this process requires expertise and manual effort to implement [3], [4]. Although LLMs demonstrate promising capabilities in code understanding, their application to deobfuscation is still limited and not fully reliable [1]. This paper explores whether publicly accessible LLMs-based chatbots can assist in deobfuscating source code and how effective they are in doing so. Our goal is to explore the practical limits of this approach and assess the potential to support developers in detecting hidden code, even without deep domain knowledge in LLMs. With findings showing that, while current chatbots still struggle with this process they can better support it when provided with contextual information and example-driven prompts.

To be disclosed in final version.

II. Proposed Method

A. Problem Definition

Code obfuscation is the process of intentionally making source code difficult to understand while preserving its functionality. It is often used for software protection or to hide malicious behaviour in malware development [3]. With the development of LLMs embedded chatbots, a new possibility emerges of assisting in reversing the deobfuscation process that typically requires expert knowledge and manual analysis of the obfuscated code. This work investigates whether publicly available chatbots can support this task effectively and under what conditions they are successful.

B. Test Design

To evaluate the performance of the chatbot Python code was obfuscated at three difficulty levels: easy, medium, and hard. A consistent prompt instructed the chatbot to act as a cybersecurity and reverse engineering expert, performing static code analysis and producing a readable, commented version of the obfuscated code completed with an explanation of the deobfuscation process [1]. In the first approach, chatbots were progressively tested with an easy-level to hard-level obfuscation code within the same chat session. In the next phase, chatbots were presented with an isolated medium-level version of a new obfuscated code first and with a hard version of yet another code in the same condition after that. Followed by a medium-level plus a hard-level version of an obfuscated code together without the easy version. On the second approach, a medium and hard level obfuscated code was introduced in isolated chat sessions, without prior examples or history. In the last phase, the process was repeated but this time, including an example of deobfuscation in the prompt. To account for possible recognition of common patterns, the tests included both well-known algorithms (e.g., prime numbers and Fibonacci) and less predictable code (e.g., visual pattern generators) [4].

C. Tools and Chatbots Used

The tests focused on five publicly accessible and free-to- use chatbots: ChatGPT, Gemini, Claude, Perplexity, and DeepSeek. These chatbots were selected from a broader set based on a preliminary interaction. The final selection benefited chatbots that demonstrated at least some capability in handling obfuscated code (e.g., gives a deobfuscation example), provided explanations on its limitations or on the process

needed to work on the deobfuscation, did not impose strict ethical restrictions, and had a significant presence or popularity among general users.

D. Evaluation Criteria

The deobfuscation performance of each chatbot was evaluated based on a set of key criteria together with a comparison with the source code before it was obfuscated in each test. The criteria were divided on a scale from 0 (lowest performance) to 5 (highest performance), each reflecting a specific aspect of the deobfuscation process. The chatbots received a total score ranging from 0 to 50 for each test, enabling a clear and consistent comparison of their performance between all tests and between them. Each chatbot performance was assessed through the following metrics:

- Obfuscation level: Assesses whether the code was completely deobfuscated and fully restored.
- Code structure preservation: Assesses the preservation of the original code after deobfuscation.
- Readability and comprehensibility: Assesses how comprehensible the deobfuscated code is.
- Deobfuscation accuracy: Assesses how accurately the obfuscation was removed without unnecessary modifications.
- Code commenting: Assesses whether the deobfuscated code includes explanatory comments.
- Deobfuscation explanation: Assesses how thoroughly the chatbot explains the deobfuscation process.
- Code executability: Assesses if the deobfuscated code is fully executable.
- Original purpose retention: Assesses whether the deobfuscated code retains the original functionality.
- Programming best practices: Assesses if the deobfuscated code follows programming best practices.
- Maintainability: Assesses the ease of maintaining and modifying the deobfuscated code.

An example of the results of the precedent criteria can be seen in Fig. 1.

E. Limitations of the Methodology

This study presents limitations that come from security restrictions and variation in results [1]. The chatbots rely on static code analysis due to security restrictions limiting their ability to detect dynamic behaviors. In addition, a variation in the results can occur between different sessions. The effectiveness of the deobfuscation process is also affected by the prompt quality and inconsistency, making it less reproducible without careful, consistent prompt engineering [2], [4].

III. Results and Discussion

In progressive tests within a single chat session (easy to hard-level obfuscations), all chatbots performed successfully, benefiting from accumulated prior history. Without history, medium-level results were generally good for known algorithms but weaker for more unfamiliar ones, and hard-level versions were mostly unsuccessful in both cases. When medium and hard versions were presented together performance slightly improved. In isolated chats with no prior history, only medium-level code was partially deobfuscated while hard-level code deobfuscation was unsuccessful. Adding an example to the prompt improved results for medium-level obfuscated code but had little to no effect on hard-level results. Perplexity uniquely handled, though not entirely successfully, a hard-level test without prior context in a new chat session, but failed with prior history. DeepSeek was the most consistent overall. The results show that context and example-driven prompts significantly influence the deobfuscation success.

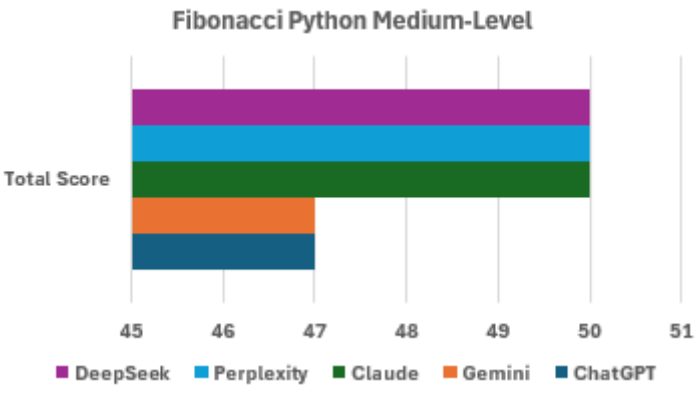


Fig. 1. Results of the deobfuscation process of a medium-level Python code without previous context.

IV. Conclusion

This study evaluated the ability of publicly accessible chatbots deobfuscating Python code. Whereas chatbots performed well with easy and some medium-level obfuscation code, especially when examples of deobfuscation where available, they consistently failed to handle harder cases without examples or history. The results suggest that prompt engineering and code familiarity (e.g., known algorithms) play a key role in the process of deobfuscation success. Although current LLMs embedded chatbots can support developers in the analysis of obfuscated code, they remain limited for advanced cases but can improve in the near future.

References

[1] C. Patsakis, F. Casino, and N. Lykousas, "Assessing llms in malicious code deobfuscation of real-world malware campaigns," *Expert Systems with Applications*, vol. 256, p. 124912, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417424017792

[2] M. A. Ferrag, F. Alwahedi, A. Battah, B. Cherif, A. Mechri, N. Tihanyi, T. Bisztray, and M. Debbah, "Generative ai in cybersecurity: A comprehensive review of llm applications and vulnerabilities," *Internet of Things and Cyber-Physical Systems*, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2667345225000082

[3] D. Lee, G. Jeon, S. Lee, and H. Cho, "Deobfuscating mobile malware for identifying concealed behaviors," *Computers, Materials and Continua*, vol. 72, no. 3, pp. 5909–5923, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1546221822009766

[4] S. Zhang, S. Li, J. Lu, and W. Yang, "Power-astnn: A deobfuscation and ast neural network enabled effective detection method for malicious powershell scripts," *Computers Security*, vol. 154, p. 104441, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167404825001300

Secure Integration of Generative AI in Video Games: Methodology, Risks, and Future Directions

Pablo Natera Muñoz
Grupo de Ingeniería de Medios (GIM)
University of Extremadura
Cáceres, Spain
pnateram@alumnos.unex.es

Pablo García Rodríguez
Grupo de Ingeniería de Medios (GIM)
University of Extremadura
Cáceres, Spain
pablogr@unex.es

María Mar Ávila Vegas
Grupo de Ingeniería de Medios (GIM)
University of Extremadura
Cáceres, Spain
mmavila@unex.es

Antonio M. Silva-Luengo
Grupo Robolab
University of Extremadura
Cáceres, Spain
agua@unex.es

Belén María Ramírez Gabardino
Grupo de Ingeniería de Medios (GIM)
University of Extremadura
Cáceres Spain
belramirez@unex.es

***Abstract*— The use of generative artificial intelligence (AI) is rapidly transforming video game development, enabling the creation of dynamic dialogues, procedural environments, and customized in-game experiences. However, this paradigm shift also introduces significant cybersecurity challenges. This paper explores the integration of generative AI in game design and the associated risks, including prompt injection, generation of harmful content, and potential data leakage. We propose a secure generative framework for video games that incorporates input sanitization, content moderation, and sandboxed model execution to mitigate these threats. Our methodology aims to balance creativity and security, enabling safe deployment of generative systems in modern game environments.**

***Keywords*— Artificial Intelligence, Generative AI, Cybersecurity, Video Games, Procedural Content Generation, Prompt Injection, Secure Game Design.**

I. Introduction

In recent years, generative artificial intelligence has emerged as a disruptive force in the video game industry. Tools capable of generating dialogues, narratives, quests, environments, and even music are reshaping the development pipeline, drastically reducing production time and enabling adaptive, player-specific experiences. Systems based on Large Language Models (LLMs) or diffusion models now allow non-playable characters (NPCs) to respond intelligently to player actions, while procedural generation powered by AI can produce entire worlds with minimal human intervention.

Despite the creative potential, the integration of generative AI in games poses a new class of cybersecurity risks. Unlike traditional deterministic systems, generative models introduce stochastic and unpredictable behaviors. This unpredictability can be exploited by malicious actors, leading to the generation of offensive content, bypassing content filters, or even manipulating the AI to gain unfair advantages [1]. Furthermore, training data used in these systems may inadvertently contain sensitive information, resulting in unintended data leakage.

Prompt injection attacks (where users craft malicious inputs to subvert the model's intended behavior) are particularly concerning in interactive settings [2]. These attacks could lead to scenarios where NPCs provide inappropriate responses or the game logic is manipulated. Additionally, the lack of robust content moderation mechanisms in real-time generation systems can result in reputational and legal risks for game developers.

This paper addresses these concerns by proposing a framework for the secure implementation of generative AI in video games. Our goal is to provide practical guidelines and a technical architecture that allows developers to harness the power of generative models while maintaining a secure and ethical gaming environment.

II. Background and related works

Generative artificial intelligence is rapidly becoming a powerful tool in video game development, enabling the creation of dynamic and adaptive content such as dialogue systems, procedural environments, narrative branching, and even audio and visual elements. Technologies based on large language models and other generative systems are being integrated into games to automate creative processes and provide highly personalized experiences. Examples like AI Dungeon, Ubisoft's Ghostwriter, and independent experiments using tools like ChatGPT have shown the potential of real-time generative content to enhance immersion and replayability. However, the creative benefits of these systems come with significant challenges, particularly when used in interactive settings where user input is unpredictable and constant.

One of the most critical concerns is the possibility of prompt injection attacks, where players manipulate input text to force the AI to behave in unintended or malicious ways. This can result in the generation of offensive, biased, or even illegal content if appropriate filters are not in place. Additionally, improperly trained models can unintentionally expose fragments of their training data, posing serious risks to privacy and intellectual property. While some industries are beginning to implement safety measures such as content moderation, human feedback alignment [3], and sandboxed execution environments, the gaming sector has yet to adopt

these practices widely. This work aims to address that gap by proposing a secure framework for the integration of generative AI in video games—one that embraces the creative potential of these models while embedding cybersecurity at the core of their design and deployment.

III. Proposed Methodology

To ensure the safe integration of generative artificial intelligence into video game environments, we propose a methodology that prioritizes cybersecurity without sacrificing creativity or interactivity. Our approach is based on the idea that any system responsible for generating content dynamically must be surrounded by protective mechanisms that control both the input provided by players and the output produced by the model. This framework does not rely on a rigid set of procedural steps but rather on a set of interconnected principles aimed at reducing potential abuse and maintaining the integrity of the player experience.

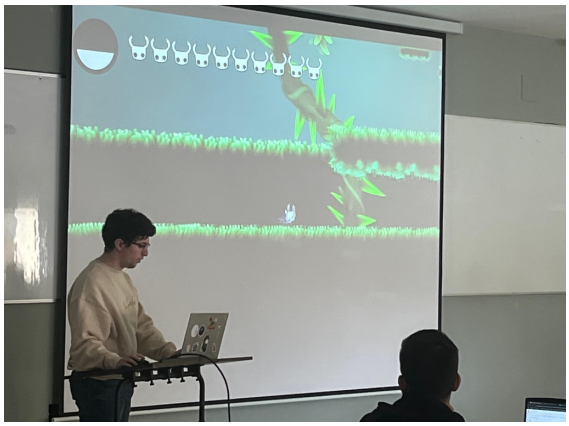


Fig. 1. Integrating AI and cyber secure properties in video games

The first key element of the methodology is input control. Since generative models are highly sensitive to prompts, it is essential to monitor and sanitize user inputs before they reach the AI system. This involves basic filtering of harmful language, but also more advanced context-aware checks that detect attempts to manipulate or trick the model through indirect instructions. By managing how players interact with the system, developers can reduce the likelihood of prompt injection attacks or unintended behavior caused by ambiguous phrasing.

Equally important is the validation of generated content. All outputs from the model—whether textual, visual, or otherwise—should be evaluated before being displayed or used within the game. This evaluation can be done through a combination of lightweight content classifiers, toxicity filters, and fallback systems that trigger re-generation or replace unsafe results with predefined alternatives. Instead of blocking creativity, these filters serve as a safety net, ensuring that generated content aligns with ethical and design standards.

Finally, the methodology encourages the isolation of the generative AI component from the rest of the game engine. By deploying the model within a sandboxed environment, developers can prevent it from accessing sensitive resources or triggering unintended side effects in gameplay [4]. Logging and monitoring mechanisms also play a crucial role, allowing teams to detect misuse, fine-tune moderation strategies, and continuously adapt the system to evolving threats [5]. While the proposed framework is flexible, it establishes a solid foundation for deploying generative AI in games responsibly and securely.

IV. Future goals

As generative AI continues to evolve, future work in this area should focus on making secure content generation more scalable, accessible, and adaptable to different game genres and platforms. One of the main priorities is the development of lightweight moderation systems that can operate in real time without disrupting the gameplay experience. These systems must balance performance with safety, especially in multiplayer or open-world environments where player interaction is unpredictable [6].

Another important goal is to explore new methods for aligning generative models with game-specific ethics and narrative constraints. This includes fine-tuning models with domain-relevant datasets and designing prompt templates that limit unintended outputs without overly restricting creativity. Additionally, we aim to improve transparency by incorporating explainable AI mechanisms that help developers and users understand how certain content is generated and why specific outputs are filtered or modified [7].

In the long term, we envision a standard framework for secure generative AI in games that could be adapted across engines like Unity or Unreal, supported by open-source tools and community-driven best practices. Collaboration between AI researchers, game developers, and cybersecurity experts will be essential to building a responsible future for generative technologies in interactive entertainment.

Acknowledgments

This initiative is carried out within the framework of the funds of the Recovery, Transformation and Resilience Plan, financed by the European Union (Next Generation) - National Cybersecurity Institute (INCIBE) in the project C107/23 “Artificial Intelligence Applied to Cybersecurity in Critical Water and Sanitation Infrastructures”.

References

[1] G. J. Branch et al., "Evaluating the Susceptibility of Pre-Trained Language Models via Handcrafted Adversarial Examples," 5 Sep. 2022. [Online]. Available: https://arxiv.org/abs/2209.02299

[2] IBM, "What is a prompt injection attack?" [Online]. Available: https://www.ibm.com/think/topics/prompt-injection

[3] Utopia Analytics, "AI content moderation for online gaming & chats," [Online].Available:https://www.utopiaanalytics.com/ai-content-moderation-for-online-gaming-and-chat-services

[4] Fortinet,"What Is Sandboxing?" [Online].Available:https://www.fortinet.com/resources/cyberglossary/what-is-sandboxing

[5] Harvard University Information Technology, "AI Sandbox," [Online]. Available: https://www.huit.harvard.edu/ai-sandbox

[6] Lasso, "AI-powered Content Moderation for Gaming Platforms," [Online].Available:https://www.lassomoderation.com/industries/content-moderation-for-gaming/

[7] NVIDIA Developer, "Securing LLM Systems Against Prompt Injection,"[Online].Available:https://developer.nvidia.com/blog/securing-llm-systems-against-prompt-injection/

Session 2

A novel architecture for IoT security

Laura Grande-Pérez
BISITE Research Group
University of Salamanca
Salamanca, España
0009-0004-2994-8628

Jesús-Ángel Román-Gallego
BISITE Research Group
University of Salamanca
Salamanca, España
0000-0002-2058-6219

Pablo Plaza Martínez
BISITE Research Group
University of Salamanca
Salamanca, España
0000-0001-7628-7632

Manuel López Pérez
BISITE Research Group
University of Salamanca
Salamanca, España
0000-0001-7316-6026

Albano Carrera
BISITE Research Group
University of Salamanca
Salamanca, España
0000-0003-0763-6434

***Abstract*— The growing prevalence of Internet of Things (IoT) devices in daily life has led to an increase in cyber-attacks, underscoring the need for robust security measures. To address this issue, it is crucial to develop architectures capable of effectively detecting and mitigating threats while meeting both user and manufacturer expectations. This paper introduces a layered reference architecture that integrates security controls at each level, ensuring regulatory compliance and protecting critical assets. The study focuses on two main aspects: identifying key assets that pose the highest risk if compromised and designing a secure architecture that aligns with cybersecurity standards. This approach provides a comprehensive framework for building and maintaining secure IoT infrastructures.**

Keywords—IoT, cybersecurity, architecture, attack.

I. INTRODUCTION

Internet of Things (IoT) devices have revolutionized daily life by enhancing connectivity in homes, workplaces, and various industries [1]. They enable automation, improve efficiency, and provide real-time services. However, their widespread use also introduces major security risks, as they handle vast amounts of sensitive data, making them prime targets for cybercriminals [2].

The rapid expansion of IoT has led to a surge in cyberattacks, affecting both hardware and software. The absence of universal security standards and the diversity of manufacturers and communication protocols create multiple vulnerabilities. Threats such as device hijacking, data breaches, and system manipulation demonstrate the urgent need for stronger security measures. Since IoT devices are constantly connected to networks, a single compromised device can endanger an entire system [2,3].

To tackle these issues, this study proposes a layered security architecture that integrates protection at every level of an IoT system. This approach ensures compliance with regulations, identifies critical assets, and embeds security from the design phase. By offering a structured framework, it aims to reduce risks and defend against both current and future threats. Strengthening IoT security will ultimately enhance trust and support the safe expansion of these technologies.

The paper is structured as follows: Section 2 reviews existing IoT security research, Section 3 presents the proposed architecture, Section 4 discusses its implementation, and Section 5 concludes the study.

II. IOT INFRASTRUCTURES SECURITY

Several architectures are designed to secure IoT environments, which involve numerous connected devices, real-time data processing, and efficient communication. While there is no universal standard for IoT architecture layers, a common model consists of four layers: perception, application, network, and physical [3,4], as show in Figure 1.

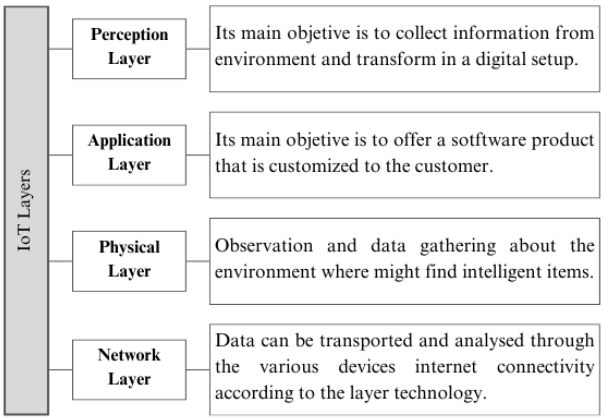


Figure 1. IoT Layers model

One notable approach is the multi-layered security architecture, which incorporates encryption, access control, and node authentication to protect IoT networks under resource constraints [5]. These layers secure different aspects such as sensing, communication, and data storage. Some proposals emphasize the IoT gateway as a critical component, particularly in smart city applications [6]. These architectures enhance security, scalability, and interoperability by enabling secure data transfers, device authentication, and access point functionalities.

Other architectures focus on networking protocols and wireless communication principles to support diverse and widespread IoT devices. End-to-end security frameworks address IoT complexities, ensuring protection from device interaction at the network edge to cloud-based applications [7]. Sustainable security models prioritize energy-efficient security mechanisms by optimizing algorithms to reduce computational demand and minimize power consumption [5].

The integration of cloud and edge computing strengthens IoT security by enabling faster data processing near the source, facilitating real-time threat detection and response. Cloud solutions provide scalable and continuously updated security measures [5]. Additionally, IoT security has evolved across three generations: traditional IT security, current widespread IoT networks, and future secure IoT ecosystems that could replace smartphones with enhanced security compliance [8].

Innovative security technologies include blockchain, which strengthens authentication and peer-to-peer communications [9]. Functional security standards, such as ISO 30141:2018, have demonstrated effectiveness in reducing cybersecurity risks [10]. Consumer-level IoT security solutions recommend trusted platform modules (TPMs) and embedded technologies [11]. However, IoT security remains complex due to its scale and heterogeneity, necessitating adaptive, context-aware approaches [12].

Artificial Intelligence (AI) and Machine Learning (ML) are increasingly used to detect and mitigate cyber threats. Solutions such as FusionNet, which integrates multiple ML models, outperform traditional anomaly detection methods [13]. Key ML techniques as reinforced, supervised, and unsupervised learning play a crucial role in identifying emerging cyberattacks and strengthening IoT defences [3].

DDoS attacks pose a serious threat to IoT infrastructure, with AI-based detection models offering promising mitigation strategies [22]. However, integrating AI solutions into heterogeneous IoT systems remains challenging. Despite advancements, DDoS protection remains insufficient in current IoT architectures [20, 23]. Privacy risks also persist due to the vast amount of sensitive data collected by IoT devices, making them vulnerable to breaches [21].

Common cyber threats such as TCP-SYN flood attacks can disrupt IoT operations, emphasizing the need for accurate risk assessments [19]. Penetration testing is widely recommended to identify system vulnerabilities and prevent security breaches [19]. Frameworks like NIST's cybersecurity framework have been successfully implemented in IoT environments, alongside real-time security assessment models [16 - 18].

To ensure robust IoT security, a comprehensive, adaptive, and continuously updated approach is required [5]. The evolution of IoT security has shifted from reactive, device-centric measures to proactive, integrated strategies that incorporate blockchain, AI, and standardized frameworks. Advanced detection and mitigation methods, continuous risk assessment, and rigorous data privacy measures are essential for the future of IoT security [6].

III. Proposed IoT Architecture

A structured and adaptable security architecture for IoT must evolve to meet emerging threats while ensuring system integrity. The proposed model that is showed in Figure 2 consists of four key layers: Perception, Network, Middleware, and Application, each integrating security mechanisms to safeguard devices and data. The use of cloud computing enables large-scale data storage and processing, while edge computing ensures localized data handling, reducing exposure to external threats.

The layers that are proposed and the security measures that are designed to protect them are described below.

A. Perception (Physical) Layer

This foundational layer comprises IoT devices (e.g., sensors, actuators) and IoT gateways, responsible for collecting and transmitting data. The gateways serve as intermediaries, connecting devices to networks or the internet.

The specific security measures designed to protect the devices and data handled by the IoT system in this layer are:

- **Device Authentication:** Enforces strong authentication mechanisms such as public key infrastructure (PKI) and digital certificates to ensure only authorized devices access the network.
- **On-Device Encryption:** Implements lightweight encryption to protect data at the source, considering IoT devices' resource constraints.
- **Firmware Security:** Uses digital signatures to verify firmware integrity and secure update mechanisms to prevent malware installations.
- **Tamper Detection:** Employs sensors to identify physical tampering attempts.

B. Network Layer

This layer facilitates data transmission between IoT devices and processing systems, relying on communication protocols and network infrastructure.

The specific security measures designed to protect the devices and data handled by the IoT system in this layer are:

- **Secure Communication:** Uses protocols such as TLS/DTLS for encrypted data exchange between devices and servers.
- **Virtual Private Networks (VPN) & Network Segmentation:** Implements VPNs and network segmentation to restrict unauthorized access and limit the attack surface.
- **Software-Defined Networking (SDN)**: Allows for dynamic security policy updates in response to detected threats.
- **Intrusion Detection and Prevention (IDS/IDP):** Employs firewalls and IDS/IDP systems to detect and mitigate malicious activities in real time.

C. Middleware Layer

Acting as an intermediary, this layer manages data processing, device management, and communication services. Cloud integration enables large-scale storage and processing, while edge computing ensures critical data is handled locally to minimize exposure. AI enhances security by detecting anomalies and predicting threats.

The specific security measures designed to protect the devices and data handled by the IoT system in this layer are:

- **Identity & Access Management (IAM):** Implements IAM solutions, including multi-factor authentication (MFA), to regulate access to resources.
- **Data Encryption (At Rest and In Transit):** Ensures encryption for stored and transmitted data to protect against unauthorized access.

- **Vulnerability Assessment and Patch Management:** Conducts regular assessments and applies timely security updates.
- **AI-Driven Security Analysis:** Uses AI algorithms to analyze traffic patterns, detect anomalies, and automate incident responses.

D. Application Layer

This layer consists of IoT applications that process data and deliver services to end users.

The specific security measures designed to protect the devices and data handled by the IoT system in this layer are:

- **Role-Based Access Control (RBAC):** Defines access permissions based on user roles, enforcing a least privilege principle.
- **Auditing and Monitoring:** Logs and monitors system activity to detect anomalies and unauthorized access.
- **Application Security:** Adopts secure coding practices and penetration testing to prevent vulnerabilities.
- **Data Encryption for Stored Data:** Protects application-stored information through encryption to ensure privacy.
- **Regulatory Compliance:** Aligns with security standards like General Data Protection Regulation (GDPR) to ensure proper handling of personal data.

E. Cross-Cutting Security Measures

Beyond individual layers, holistic security mechanisms are essential for continuous protection across the IoT system:

- **Security Information and Event Management (SIEM):** Real-time security monitoring across layers for rapid incident detection and response.
- **Log Collection Agents:** Aggregates security logs for analysis.
- **Event Correlation and AI Analysis:** Uses AI-driven analysis to identify advanced threats.
- **Automated Incident Response:** Triggers actions such as device isolation upon detecting known attack patterns.
- **Cyber Resilience Framework:** Establishes protocols for threat detection, response, recovery, and continuous improvement in cybersecurity.

This layered IoT security architecture adopts a security-by-design and privacy-by-design approach to address IoT's complex security challenges. By integrating security mechanisms at each layer and ensuring seamless collaboration, the model enhances protection against emerging threats, fostering a resilient and secure IoT ecosystem.

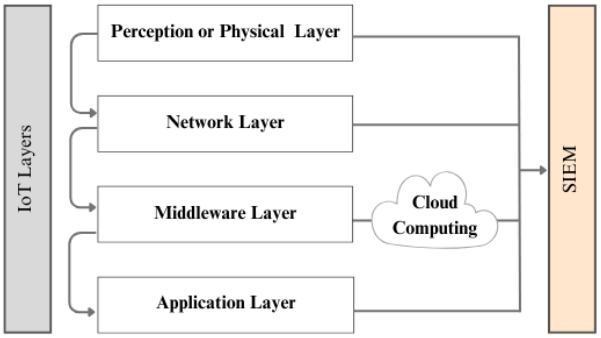


Figure 2. Proposed Architecture Diagram

IV. Example of Implementation

To illustrate the potential and versatility of the proposed architecture, a real use case is presented where its implementation addresses specific cybersecurity challenges.

Healthcare: Remote Patient Monitoring

Healthcare systems increasingly rely on IoT devices for remote patient monitoring (RPM), enabling healthcare professionals to track patients' health conditions beyond traditional clinical settings. This approach enhances proactive care, facilitates early detection of health issues, and reduces the need for frequent hospital visits. However, implementing secure IoT architectures in healthcare is crucial to protect sensitive patient data and ensure compliance with stringent regulatory requirements.

The proposed layered security architecture for remote patient monitoring incorporates robust security mechanisms at each level to mitigate cyber threats and ensure the integrity, confidentiality, and availability of patient data.

Each layer is listed below with the security mechanisms that apply to this case.

A. Perception (Physical) Layer

- Lightweight encryption techniques are applied to safeguard health data collected by wearable devices and medical sensors.
- Tamper detection mechanisms prevent unauthorized access or alterations to IoT-enabled healthcare devices.
- Secure device authentication ensures that only trusted medical IoT devices transmit data.

B. Network Layer

- Secure communication protocols (e.g., TLS, DTLS) are employed to encrypt health data transmissions, preventing interception and ensuring the confidentiality of sensitive patient information.
- VPNs and network segmentation limit unauthorized access, isolating critical healthcare devices from potential cyber threats.
- IDS/IDP monitor the network to identify and mitigate malicious activities targeting patient data.

C. Middleware Layer

- IAM systems enforce strict access controls, allowing only authorized personnel to access and process

patient data while maintaining privacy-by-design principles.

- Edge computing enables real-time processing of critical patient data at the device level, reducing latency and minimizing exposure to cyber risks.
- Cloud-based solutions facilitate large-scale storage and advanced analytics while ensuring compliance with regulatory frameworks.
- AI-driven security mechanisms analyse patterns in patient data to detect anomalies that may indicate security breaches or potential health issues.

D. Application Layer

- Secure health monitoring applications are developed with built-in encryption and security best practices to comply with regulations such as GDPR and Health Insurance Portability and Accountability Act (HIPAA).
- RBAC ensures that healthcare professionals can access only the patient data relevant to their responsibilities.
- Comprehensive audit logging and real-time monitoring track user activities to detect unauthorized access attempts and security violations.

E. Cross-Layer Security Measures

- SIEM systems continuously monitor all layers of the architecture to detect and respond to security incidents in real time.
- Periodic security audits assess vulnerabilities and ensure compliance with industry security standards.
- Cybersecurity risk management strategy includes proactive threat mitigation plans and response frameworks to handle potential breaches effectively.

Among the benefits of the application of this model, it ensures the confidentiality and integrity of health data, enhancing patient trust in digital healthcare systems and facilitating proactive care. The SIEM system continuously monitors all layers of the architecture to ensure a rapid response to any security breach:

- **Data confidentiality and integrity:** The layered security approach ensures that sensitive patient information remains protected from unauthorized access and tampering.
- **Regulatory compliance:** Adherence to GDPR, HIPAA, and other healthcare security regulations builds patient trust and avoids legal liabilities.
- **Enhanced patient confidence:** Robust security mechanisms reinforce trust in digital healthcare systems, encouraging wider adoption of remote monitoring solutions.
- **Proactive healthcare:** The real-time monitoring of patients' health conditions allows early detection of medical issues, leading to faster interventions and improved patient outcomes.
- **Rapid incident response:** The SIEM system provides continuous security oversight across all layers, enabling immediate response to potential breaches and cyber threats.

This secure IoT-based remote patient monitoring architecture establishes a comprehensive and resilient cybersecurity framework, ensuring privacy, security, and efficiency in digital healthcare environments.

V. Conclusions

The implementation of a four-layered IoT security architecture represents a structured, adaptable, and comprehensive approach to mitigating the complex security risks inherent in IoT ecosystems. By integrating security-by-design principles, this architecture ensures that protection mechanisms are embedded from the initial stages of development, minimizing vulnerabilities across perception, network, middleware, and application layers. The adoption of lightweight encryption techniques, secure authentication protocols, and firmware integrity verification at the device level enhances the security of IoT endpoints, preventing unauthorized access and potential system compromise. Furthermore, secure data transmission mechanisms, such as TLS/DTLS encryption, VPNs, and network segmentation, safeguard the confidentiality and integrity of IoT-generated data during transit. A layered security model also allows organizations to enforce identity and access control policies, ensuring that only authorized users and devices can interact with sensitive IoT resources. In this context, IAM, MFA and RBAC play a crucial role in strengthening data privacy and preventing unauthorized exposure.

Beyond conventional security measures, AI and ML-driven anomaly detection enhance predictive security capabilities, allowing for the early identification of potential cyber threats. AI-powered security analytics improve incident detection, automated response, and risk mitigation, ensuring proactive defence mechanisms against evolving attacks. Additionally, compliance with regulatory frameworks such as GDPR, HIPAA, and ISO 30141:2018 is critical in fostering trust and transparency in digital healthcare, smart cities, and industrial IoT applications. The integration of SIEM systems further strengthens IoT security by providing real-time monitoring, log correlation, and automated incident response across all layers. These elements collectively establish a resilient cybersecurity posture that not only mitigates risks but also enables the scalability, interoperability, and sustainability of IoT deployments. As IoT continues to evolve, organizations must remain committed to continuous security enhancements, adaptive threat intelligence, and holistic risk management strategies to ensure the long-term reliability and safety of connected environments.

Acknowledgment

This activity is conducted as part of the Strategic Project Secure Certified Resources in IoT Networks (SCRIN) project (C068/23), established through a collaboration agreement between the National Institute of Cybersecurity (INCIBE) and the University of Salamanca. It falls within the framework of the Recovery, Transformation, and Resilience Plan, funded by the European Union (Next Generation). This Spanish government initiative serves as a roadmap for modernizing the economy, fostering economic growth, and creating jobs, ensuring a robust, inclusive, and resilient recovery from the COVID-19 crisis while addressing the challenges of the coming decade.

REFERENCES

[1] Ojo, M. O., Giordano, S., Procissi, G., & Seitanidis, I. N. (2018). A review of low-end, middle-end, and high-end IoT devices. IEEE Access, 6, 70528-70554.

[2] Haroon, A., Shah, M. A., Asim, Y., Naeem, W., Kamran, M., & Javaid, Q. (2016). Constraints in the IoT: the world in 2020 and beyond. International Journal of Advanced Computer Science and Applications, 7(11).

[3] M. Z. Khan and M. U. Bokhari, "Comparative Study of Detection of IoT Attacks using Machine Learning Techniques," 2023 10th International Conference on Computing for Sustain-able Global Development (INDIACom), New Delhi, India, 2023, pp. 1648-1651.

[4] A. Munshi, N. A. Alqarni and N. Abdullah Almalki, "DDOS Attack on IOT Devices," 2020 3rd International Conference on Computer Applications & Information Security (ICCAIS), Riyadh, Saudi Arabia, 2020, pp. 1-5, doi: 10.1109/ICCAIS48893.2020.9096818.

[5] J. Zhang, H. Jin, L. Gong, J. Cao and Z. Gu, "Overview of IoT Security Architecture," 2019 IEEE Fourth International Conference on Data Science in Cyberspace (DSC), Hangzhou, China, 2019, pp. 338-345, doi: 10.1109/DSC.2019.00058

[6] H. Y. Ali and W. El-Medany, "IoT security: A review of cybersecurity architecture and layers," 2nd Smart Cities Symposium (SCS 2019), Bahrain, Bahrain, 2019, pp. 1-7, doi: 10.1049/cp.2019.0191

[7] R. -A. Craciun, R. N. Pietraru and M. A. Moisescu, "Secure IoT Gateway: The First Layer of Cybersecurity for Secure Infrastructure in Smart Cities," 2023 IEEE International Smart Cities Conference (ISC2), Bucharest, Romania, 2023, pp. 1-5, doi: 10.1109/ISC257844.2023.10293352.

[8] C. Vorakulpipat, E. Rattanalerdnusorn, P. Thaenkaew and H. Dang Hai, "Recent challenges, trends, and concerns related to IoT security: An evolutionary study," 2018 20th International Conference on Advanced Communication Technology (ICACT), Chuncheon, Korea (South), 2018, pp. 405-410, doi: 10.23919/ICACT.2018.8323774.

[9] A. K. Singh and N. Kushwaha, "Software and Hardware Security of IoT," 2021 IEEE Inter-national IOT, Electronics and Mechatronics Conference (IEMTRONICS), Toronto, ON, Canada, 2021, pp. 1-5, doi: 10.1109/IEMTRONICS52119.2021.9422651.

[10] G. Gómez, E. Espina, J. Armas-Aguirre and J. M. M. Molina, "Cybersecurity architecture functional model for cyber risk reduction in IoT based wearable devices," 2021 Congreso Internacional de Innovación y Tendencias en Ingeniería (CONIITI), Bogotá, Colombia, 2021, pp. 1-4, doi: 10.1109/CONIITI53815.2021.9619624.

[11] S. J. Johnston, M. Scott and S. J. Cox, "Recommendations for securing Internet of Things devices using commodity hardware," 2016 IEEE 3rd World Forum on Internet of Things (WF-IoT), Reston, VA, USA, 2016, pp. 307-310, doi: 10.1109/WF-IoT.2016.7845410.

[12] S. K. Rajput, R. Umamageswari, R. Singh, L. Thakur, C. P. Sanjay and M. K. Chakravarthi, "Understanding IoT Security. Chakravarthi, "Understanding IoT Security: A Point-by-point Investigation of IoT Weaknesses and a First Exact Glance at Web Scale IoT Exploits," 2022 5th International Conference on Contemporary Computing and Informatics (IC3I), Uttar Pradesh, India, 2022, pp. 1677-1683, doi: 10.1109/IC3I56241.2022.10073056.

[13] D. Alsalman, "A Comparative Study of Anomaly Detection Techniques for IoT Security Using Adaptive Machine Learning for IoT Threats," in IEEE Access, vol. 12, pp. 14719-14730, 2024, doi: 10.1109/ACCESS.2024.3359033.

[14] Z. Mohammad, T. A. Qattam and K. Saleh, "Security Weaknesses and Attacks on the Inter-net of Things Applications," 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), Amman, Jordan, 2019, pp. 431-436, doi: 10.1109/JEEIT.2019.8717411.

[15] N. Neshenko, E. Bou-Harb, J. Crichigno, G. Kaddoum and N. Ghani, "Demystifying IoT Security: An Exhaustive Survey on IoT Vulnerabilities and a First Empirical Look on Internet-Scale IoT Exploitations," in IEEE Communications Surveys & Tutorials, vol. 21, no. 3, pp. 2702-2733, third quarter 2019, doi: 10.1109/COMST.2019.2910750.

[16] J. Webb and D. Hume, "Campus IoT collaboration and governance using the NIST cybersecurity framework," Living in the Internet of Things: Cybersecurity of the IoT - 2018, Lon-don, 2018, pp. 1-7, doi: 10.1049/cp.2018.0025.

[17] S. Ribeiro, J. P. De Lima Cassiano and A. Almeida, "CSecPrivAF - Cybersecurity, and Privacy Assessment Framework for IoT Systems," 2022 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 2022, pp. 1094-1095, doi: 10.1109/CSCI58124.2022.00195.

[18] S. K. Datta, "DRAFT - A Cybersecurity Framework for IoT Platforms," 2020 Zooming In-novation in Consumer Technologies Conference (ZINC), Novi Sad, Serbia, 2020, pp. 77-81, doi: 10.1109/ZINC50678.2020.9161441.

[19] European Commission, Directorate-General for Communications Networks, Content and Technology, (2022). Cyber resilience act: new EU cybersecurity rules ensure more secure hardware and software products, European Commission. https://data.europa.eu/doi/10.2759/543836

[20] S. Kumar, A. Guerrero and C. Navarro, "Cyber Security Flood Attacks and Risk Assessment for Internet of Things (IoT) Distributed Systems," 2023 IEEE World AI IoT Congress (AIIoT), Seattle, WA, USA, 2023, pp. 0392-0397, doi: 10.1109/AIIoT58121.2023.

[21] D. K. Alferidah and N. Jhanjhi, "Cybersecurity Impact over Bigdata and IoT Growth," 2020 International Conference on Computational Intelligence (ICCI), Bandar Seri Iskandar, Malaysia, 2020, pp. 103-108, doi: 10.1109/ICCI51257.2020.9247722.

[22] N. Zainuddin, M. Daud, S. Ahmad, M. Maslizan and S. A. L. Abdullah, "A Study on Pri-vacy Issues in Internet of Things (IoT)," 2021 IEEE 5th International Conference on Cryptography, Security and Privacy (CSP), Zhuhai, China, 2021, pp. 96-100, doi: 10.1109/CSP51677.2021.9357592.

[23] A. Munshi, N. A. Alqarni and N. Abdullah Almalki, "DDOS Attack on IOT Devices," 2020 3rd International Conference on Computer Applications & Information Security (ICCAIS), Riyadh, Saudi Arabia, 2020, pp. 1-5, doi: 10.1109/ICCAIS48893.2020.9096818.

Trust Seal for Cybersecurity Compliance in the CENCYL Region

1st Manal Jammal
Air Institute
Salamanca, Spain
ORCID: 0009-0002-9064-7360

2nd Pedro Pinto
Guarda Polytechnic Institute
Guarda, Portugal
ORCID: 0009-0008-1543-7009

3rd Javier Parra-Domínguez
University of Salamanca
Salamanca, Spain
ORCID: 0000-0002-1088-9152

4th Juan José-González
Supercomputación Castilla y León (SCAYLE)
Spain
ORCID: 0009-0009-8126-4393

5th Pedro R. M. Inácio
University of Beira Interior
Covilhã, Portugal
ORCID: 0000-0001-8221-0666

***Abstract*—The Trust Seal is a certification framework designed to ensure cybersecurity compliance among small and medium-sized enterprises (SMEs) and public entities in the cross-border CENCYL region (Castilla y León-Central Portugal). The initiative addresses the lack of standardization and insufficient cybersecurity capacity that limit the secure adoption of digital technologies. The Trust Seal model evaluates digital solutions across three core dimensions: cybersecurity, security by design, and legal compliance with frameworks such as GDPR, NIS2, ENS, and the Cyber Resilience Act (CRA). The platform leverages a weighted scoring system to classify services under Bronze, Silver, or Gold tiers. Certificates are digitally issued and recorded on a blockchain using smart contracts, ensuring traceability, authenticity, immutability, and resistance to fraud. The solution adopts a modular microservices architecture and provides an accessible interface tailored for organizations with limited cybersecurity expertise. This initiative contributes to a safer and more trustworthy digital ecosystem and fosters regional digital resilience.**

***Index Terms*—Cybersecurity, Trust Seal, Compliance, SMEs, Blockchain, Smart Contracts, Digital Certification, CENCIL Region.**

I. Introduction

The creation of the Trust Seal arises from the need to ensure the security and reliability of digital solutions in the CENCYL region (Castilla y León – Central Portugal), especially for SMEs and public bodies with limited resources and knowledge in cybersecurity. The absence of clear standards hinders the adoption of secure digital solutions, generating distrust in the market. In the digital age, trust in information resources and technology can be diminished. We are often told that information requires higher standards of verification in digital environments than in 2 the paper world. Similarly, when we encounter digital records and files, we may have uncertainty about how much we can trust them [1], [2]. Digital trust has become a crucial element in the e-commerce landscape and the adoption of digital technologies. Consumers and businesses alike need to be confident in the security of their online interactions and in the protection of their personal and financial information. Growing security concerns in the digital realm underscore the importance of addressing users' perceived security to encourage adoption of technologies such as blockchain. Perceived security, defined as the extent to which a user considers it safe to perform activities in certain contexts, significantly influences the decision to adopt or reject technological services [3]. In the context of e-commerce, trust, perceived risk and security are central factors that influence online purchasing behavior. Consumers rely heavily on trust that their personal and financial information is protected from cyber-attacks [4]. The creation of the Trustmark seeks to directly address this need for security and trust in digital solutions for SMEs and public agencies. The Trustmark initiative aligns with the trend in computer science and archival studies that recognize the importance of provenance and context in assessing the trustworthiness of data and digital records [1]. While data should be reusable, each piece of information should carry evidence of its original history and contexts to help those who find it judge its reliability. The seal certifies that a digital solution complies with European and national standards (Cybersecurity Act, NIS2, ENS, GDPR), facilitating the adoption of secure technologies and increasing the competitiveness of companies. Certification and registration in the Marketplace will enable certified companies to gain greater visibility and credibility, encouraging continuous improvement and the adoption of sound practices in technological development. The adoption of clear standards, such as those certified by the Trust Seal, is fundamental to overcome distrust and foster a safer and more reliable digital environment, especially for those with limited resources in cybersecurity [5]. Although guidelines for digitization exist, they are often intended for production and are not easily understood by end users or libraries that rely on digital copies. The Trustmark, in this sense, seeks to provide a consumer-oriented standard that denotes that a digital solution meets relevant security and reliability criteria, thus facilitating adoption and building trust in the regional digital marketplace [6], [7]. This article is structured as follows: Section 2 presents the methodology and conceptualization of the Trust Seal,

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

including the definition of its key evaluation dimensions, the scoring system, and the development of the platform. Section 3 describes the implementation, detailing the use of blockchain technology, smart contracts, and the operational flow of the certification process. Finally, Section 4 presents the conclusions.

II. Methodology and Conceptualization

The Trust Seal certification model was designed with accessibility in mind, particularly for SMEs and public entities with limited experience in cybersecurity. Its evaluation framework is based on a structured questionnaire aligned with widely recognized standards and legal requirements, articulated through three core dimensions: cybersecurity, security by design, and compliance with GDPR and related frameworks.

A. Evaluate Dimensions

- Cybersecurity: This dimension evaluates the solution's ability to protect itself against digital threats. It includes criteria such as system integrity, incident detection and response, identity and access management, continuous monitoring, vulnerability management, and mitigation protocols. These aspects are aligned with the ISO/IEC 27001:2022 standard and the Cyber Resilience Act (CRA), which emphasizes secure configurations and proactive risk management throughout the product lifecycle [10].
- Security by Design: This dimension verifies the integration of security measures from the earliest stages of development. It covers secure development practices, dependency management, penetration testing, and the application of defensive coding techniques. This approach follows the security by design principles proposed by ENISA and promoted in the NIS2 Directive [11].
- Regulatory Compliance and GDPR: This dimension evaluates whether the solution complies with applicable legal frameworks such as the GDPR, the NIS2 Directive, the ENS, and the CRA. Key indicators include data minimization, encryption, information lifecycle management, protection of user rights, and lawful processing. For example, the ENS establishes requirements for traceability, access control, and formal incident response procedures for both the public and private sectors in Spain [9].

These dimensions were defined after a systematic review of the relevant legal texts and technical standards. From this review, a set of common principles such as data encryption, traceability, incident response, and access control was identified and translated into assessable subcategories.Each subcategory is linked to one or more specific articles of the analyzed regulations.

B. Regulatory alignment

The trust seal is based on current European legislation and international standards, with the aim of ensuring both regulatory compliance and interoperability. Instead of relying on arbitrary requirements, the model integrates legal obligations and technical guidelines recognized throughout the European Union. The following frameworks form the foundation of the certification process: the CRA, which defines cybersecurity requirements for the entire lifecycle of digital products, emphasizing secure configurations, vulnerability management, and continuous risk assessment; the NIS2 Directive, which expands cybersecurity obligations to a broader set of essential entities, addressing risk management, technical controls, and incident notification; the GDPR, together with its national implementations, which establishes principles such as data minimization, encryption, and rights management; the ENS, which is used in both the public and private sectors in Spain, and sets out requirements for access control, traceability, and incident response; and the ISO/IEC 27001:2022 standard, which provides a risk-based framework for information security management and continuous improvement. This regulatory and technical alignment ensures that certified solutions can operate securely in cross-border environments, particularly within the Iberian context.

C. Scoring Model

The evaluation process employs a weighted scoring model to assess responses to closed-form questions under each dimension. Each question is assigned a score corresponding to the maturity level of the cybersecurity practice implemented. The model uses subcategory weighting to prevent high scores in one dimension from compensating for deficiencies in another. Each dimension is scored out of 100 points. Broader subdimensions carry more weight. If the digital solution obtains a maximum score higher than 90, it is awarded the Gold Seal. If the score is higher than 80 (and below 90), the Silver Seal is awarded. If the score is higher than 75 (and below 80), the Bronze Seal is awarded. If the score does not exceed 75 points, the solution is not certified. To account for variability, questions deemed not applicable (N/A) are excluded without penalty, receiving the maximum score for fairness. The model supports flexibility, consistency, and regular updates to integrate new standards and practices.

D. Adaptability and Sector-Specific Extension

Although the trust seal has been designed to be applicable to all types of digital products and services, one of its key strengths is precisely its adaptability. From the beginning, we knew it would make no sense to evaluate a mobile application for citizen services with the same criteria as a healthcare system connected to critical infrastructure. For this reason, the model has been built in a modular and flexible way. Each block of the questionnaire can be adjusted depending on the context, allowing both the contents and the weightings of the questions to be adapted. This opens the door to different levels of specialization without the need to completely redesign the model. For example, if a solution belongs to the healthcare sector, aspects such as traceability, management of highly sensitive data, or continuous availability of the service can be reinforced. On the other hand, if it is an educational system, the protection of minors, consent management, or the secure retention of digital content may carry more weight.

Additionally, the model is ready to incorporate new regulatory requirements that may arise. We know that legal frameworks evolve as has already occurred with the entry into force of the CRA or the expansion of entities affected by NIS2 and we wanted to ensure that the trust seal does not become outdated over time. That is why the subcategories can be expanded or modified without having to redo the entire system. It is also possible to adapt the trust seal to environments outside the European framework. Thanks to the documented regulatory traceability, it would be relatively straightforward to replace legal references with their equivalents in other jurisdictions or even create local variants of the seal for specific regions. This open design of the model means it is not a rigid system, but rather a living tool that can grow and specialize based on the needs of different sectors, new regulations, and the users themselves. This ensures that the trust seal remains useful, credible, and relevant as the digital ecosystem evolves.

III. Implementation

This section outlines the technical architecture supporting the Trust Seal platform, focusing on the blockchain infrastructure, smart contract development, and operational workflow of the certification process.

A. Blockchain infrastructure

The Trust Seal leverages blockchain technology to ensure the authenticity, integrity, and traceability of issued certificates. Blockchain functions as a decentralized digital ledger, where each block contains information about transactions or events. Each block is identified by a hash that links it to the previous one, forming a chain. This technology ensures immutability, as a transaction is recorded in a new block and then added to the chain. Through a process of tokenization, signatures, and decentralized storage, this technology is well-suited for issuing and validating trust seals in the field of cybersecurity.The implementation is based on Hyperledger Besu, an open-source Ethereum client developed by the Hyperledger Foundation. Besu supports both public and private permissioned networks, making it suitable for enterprise-level decentralized applications. The platform supports Solidity smart contracts and can operate as a full Ethereum node. To achieve consensus, the platform uses the Istanbul Byzantine Fault Tolerance (IBFT) 2.0 consensus algorithm. This protocol is tailored for permissioned environments and provides Byzantine Fault Tolerance, ensuring operational reliability even when some nodes are compromised or malfunction. The algorithm guarantees deterministic finality, which means that once a block is validated, it becomes immutable.

B. Smart Contracts Architecture

The platform uses two smart contracts developed in Solidity for the Ethereum Virtual Machine (EVM): Ownable.sol and TraceDocument.sol. These contracts serve as the core of the Trust Seal's document traceability system within an EVM-compatible blockchain network. The Ownable.sol contract implements an access control mechanism based on single ownership. It restricts critical operations to the address designated as the root, with the security of the contract depending on proper private key management. The contract includes state variables, access control modifiers, event emitters, and functions to initialize and transfer ownership securely. The TraceDocument.sol contract builds on the access control of Ownable.sol and provides a registry for on-chain document traceability. It allows the contract owner to register and remove document references, identified by cryptographic hashes. Each document entry includes a unique identifier, an additional verification hash, and a validity flag. The contract also enables third parties to verify a document's presence and retrieve its identifier by providing the correct hash and verification string.

C. Contracts Interaction Logic

Interaction with deployed contracts is managed via a wrapper library. The interaction system manages the lifecycle of certification documents and allows operations such as adding, removing, and verifying document identifiers. To simplify interaction with the blockchain, the system uses the @asanrom/smart-contract-wrapper library, which abstracts the low-level details of Ethereum communication. The code establishes a connection with the Ethereum node through a remote procedure call (RPC) using HTTP. This connection is configured with a specific endpoint and timeout setting, and transactions are authenticated using a private key. Once connected, the system creates an instance of the deployed contract using its address and Application Binary Interface (ABI). Several asynchronous functions allow for contract interaction, including registering a new document, removing an existing one, and retrieving a document's identifier using its hash and verification string. These operations are wrapped in try-catch blocks to handle potential execution errors. The smart contract logic is encapsulated in the TraceDocumentWrapper class, which simplifies function calls and provides a clean interface for deployment and interaction. This class includes methods for accessing and modifying contract data, transferring ownership, and retrieving emitted events. Internally, it uses the SmartContractInterface class to encode function calls, send transactions, and decode responses. It also defines a helper class to organize events emitted by the smart contract, allowing for structured event processing. A demonstration function shows how the contract can be used in practice, by adding a document and retrieving its identifier. Notes in the code emphasize the importance of secure private key storage, error handling, and maintaining ABI compatibility.

D. Operational Workflow

The operational process of the Trust Seal platform is structured into five sequential phases that guide applicants from initial registration to digital certification. These steps are illustrated in Figure 1, which provides an overview of the operational flow used by the platform to ensure transparency, traceability, and regulatory alignment. The process begins with platform registration, where the applicant completes a form with basic organizational information. Once access is granted,

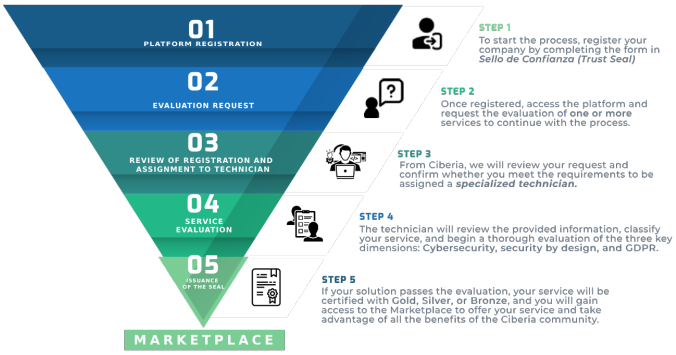


Fig. 1. Operational workflow for Trust Seal certification and registration in the CIBERIA marketplace.

the organization proceeds to the evaluation request, selecting the services to be certified and filling in a structured questionnaire that captures technical and operational data. In the third phase, an administrator requests and assigns a technician to conduct the evaluation. This ensures each case is analyzed by a qualified expert who can interpret the provided documentation and context. The service evaluation stage involves scoringthe submitted information according to the three core dimensions of the Trust Seal model: cybersecurity, security by design, and compliance. The scoring algorithm adjusts for the organization's profile and guarantees fair treatment of SMEs with limited resources. If the service meets the requirement thresholds, the system proceeds to the issuance of the seal. A digital certificate is generated, recorded on the blockchain, and published in the CIBERIA marketplace, increasing the certified service's credibility and visibility.

IV. Conclusions and Future Work

This paper presented the design and implementation of the Trust Seal, a certification framework that enhances digital trust by evaluating cybersecurity compliance in digital services offered by SMEs and public entities. Through its structured methodology, the model incorporates technical, design, and legal dimensions to ensure that certified solutions align with European regulations such as GDPR, NIS2, ENS, and CRA. By integrating blockchain and smart contract technologies, the platform guarantees certificate immutability, traceability, and resistance to fraud. The Trust Seal responds to a pressing need in the cross-border CENCYL region, where digital service providers often face limitations in technical expertise and access to standardized cybersecurity frameworks. By offering a user-friendly and legally based evaluation model, the platform contributes to strengthening digital resilience, increasing market transparency, and improving the overall perception of security in the regional digital ecosystem. Future work will focus on three main areas. First, further usability improvements to the platform will be explored based on feedback from users, including refining the questionnaire design, enhancing accessibility for non-technical profiles, and simplifying interactions with blockchain components. Second, efforts will be made to expand the geographical scope of the Trust Seal beyond the CENCYL region, adapting it for broader deployment across the Iberian Peninsula. This broader deployment would not require major adaptation, as the Trust Seal is already aligned with the regulatory frameworks of both Portugal and Spain. Third, the scope of certification may be extended to cover emerging risk areas, such as evaluating the robustness of artificial intelligence systems, or assessing post-quantum cryptographic readiness. These developments aim to keep the Trust Seal aligned with evolving digital threats and regulatory landscapes, consolidating its role as an instrument for cybersecurity governance.

Acknowledgment

This work has been partially funded by the European Regional Development Fund (FEDER) under the Interreg VI-A Spain–Portugal POCTEP 2021–2027 programme. Project:0192-CIBERIA-3-E. Also, This activity is carried out in execution of the Chair in cybersecurity (Cyberchain) C058/23. The result of a collaboration agreement signed between the National Institute of Cybersecurity (INCIBE) and the University of Salamanca. This initiative is carried out within the framework of the funds of the Recovery Plan, Transformation and Resilience Plan funds, financed by the European Union (Next Generation), the Spanish Government's roadmap for the modernization of the Spanish economy, recovery of growth and the Spanish economy, the recovery of economic growth and the creation of jobs, for a solid economic reconstruction and job creation, for solid, inclusive and resilient economic reconstruction in the wake of the COVID-19 crisis and to respond to the challenges of the next decade.

References

[1] Yeo, G. (2013). Trust and context in cyberspace. *Archives and Records*, 34(2), 214-234. https://doi.org/10.1080/23257962.2013.825207

[2] Jones, S., Wilikens, M., Morris, P., & Masera, M. (1999). Trust requirements in e-business: A conceptual framework. Technical report. University of Hertfordshire.

[3] Shin, D. D. (2019). Blockchain: The emerging technology of digital trust. *Telematics and Informatics*, 45, 101278. https://doi.org/10.1016/j.tele.2019.101278

[4] Handoyo, S. (2024). Purchasing in the digital age: A meta-analytical perspective on trust, risk, security, and e-WOM in e-commerce. *Heliyon*, 10(8), e29714. https://doi.org/10.1016/j.heliyon.2024.e29714

[5] Jacobs, J. A., & Jacobs, J. R. (2013). The digital-surrogate seal of approval: a consumer-oriented standard. *D-Lib Magazine*, 19(3).

[6] Gritzalis, S., & Gritzalis, D. (2001). A digital seal solution for deploying trust on commercial transactions. *Information Management & Computer Security*, 9(2), 71-79. https://doi.org/10.1108/09685220110388836

[7] Kluiters, L., Srivastava, M., & Tyll, L. (2023). The impact of digital trust on firm value and governance: an empirical investigation of US firms. *Society and Business Review*, 18(1), 71-103. https://doi.org/10.1108/SBR-07-2021-0119

[8] Amutio Gómez, M. A. (2016). El Esquema Nacional de Seguridad, al servicio de la ciberseguridad del sector público. *Revista Economía Industrial*, 410, 125–132.

[9] Government of Spain (2022). Royal Decree 311/2022, of May 3, which regulates the National Security Framework (ENS). Oficial State Gazette, No. 106.

[10] ISO/IEC 27001:2022. Information security, cybersecurity and privacy protection. Information security management systems. https://www.iso.org/standard/27001

[11] European Union (2024). Regulation (EU) 2024/2847 of the European Parliament and of the Council of 23 October 2024 on horizontal cybersecurity requirements for products with digital elements and amending Regulations (EU) No 168/2013 and (EU) No 2019/1020 and Directive (EU) 2020/1828. https://eur-lex.europa.eu/eli/reg/2024/2847/oj

Characterization of web server scanning tools in production environments

Ricardo Covelo*, Dimitri Silva*, João Paulo Barraca†, and João Rafael Almeida*
* IEETA / DETI, LASI, Universidade de Aveiro, Portugal
† Instituto de Telecomunicações, DETI, Universidade de Aveiro, Portugal

***Abstract*—Vulnerability scanners are critical for ensuring web application security, yet their deployment in production environments remains limited due to concerns about operational impact and infrastructure disruption. While much of the existing research primarily focuses on detection accuracy, key operational factors, such as resource consumption and scanning efficiency, are often neglected.**

This paper aims to investigate the impact of web server vulnerability testing tools, specifically evaluating open-source DAST scanners. A practical methodology is presented to evaluate scanning tools based on both security effectiveness and operational cost, allowing organizations to make informed decisions that conform to their environment requirements. The evaluation reveals significant variations in resource consumption and scanning efficiency between different tools, with measurable differences in their impact on target system performance.

***Index Terms*—DAST, Vulnerabilty Scanning, Automated Penetration Testing, Web Vulnerabilities.**

I. Introduction

More than half of organizations have reported an increase in attacks each year in the last six years, with 55% seeing an increase in 2023 [1]. Despite ongoing advancements in the cybersecurity field, namely tools and regulations, many specialists still lack confidence in detecting and responding to cyber threats. To stay ahead of these risks, organizations must proactively test and apply strong cybersecurity practices [2]. A key component of this is the integration of security testing into the application development lifecycle to detect vulnerabilities early and prioritize them for an efficient and cost-effective mitigation. However, experienced professionals are expensive, scarce and prone to human error, making it challenging to maintain a reliable method for identifying critical flaws in software design and code.

Vulnerability scanners can automate and streamline the process while also being cost-effective. These tools are configurable, can be scheduled, and can be integrated into CI/CD pipelines, improving security throughout the development lifecycle, and enabling earlier detection of vulnerabilities [3]. They can also feed valuable data into systems like Security Information and Event Managements (SIEMs) or defect management platforms. With the many advantages over traditional testing these tools provide, these also come with serious issues: false alerts, unrecognized vulnerabilities, and high resource consumption remain significant barriers to the adoption of these technologies. These issues are further amplified by the prevalent one-size-fits-all approach, which limits the effectiveness of scanners in many organizations. The characteristics of these tools, such as resource consumption, speed, extent, and precision, may not align with the specific needs and environment of the organization, leading to suboptimal results [4] or lack of adoption.

These challenges are magnified by the lack of awareness regarding scanner operational profiles, potentially leading to suboptimal tool selection. Organizations without clear understanding of performance characteristics risk deploying solutions incompatible with their existing infrastructure requirements, resulting in both underutilized capabilities and diminished trust in automated security methodologies. This underline the necessity for tool evaluations that enable practitioners to map scanner attributes—including resource demands, scanning velocity, and detection capabilities—to specific operational constraints.

Although this problem is somewhat addressed by several studies that provide valuable comparisons of different scanners, they often overlook metrics that may be even more critical to some environments than the commonly used True Positive Rate (TPR) (which measures the scanner's ability to detect vulnerabilities) and False Positive Rate (FPR) (which measures the proportion of identified vulnerabilities that are actual threats), such as speed and resource consumption in the target system [3], [5], [6].

This study aims to quantify the operational impact of Open-Source DAST tools through systematic benchmarking, providing actionable information for tool selection.

This work is organized as follows. Section I presents the introduction to the work presented, while Section II presents background knowledge regarding the context, use cases and applications. Section III describes the tools analysed and methods, while Section IV the results obtained. Finally, we draw our main conclusions in Section V.

II. Background

An effective, systematic and widely adopted way to communicate information about system vulnerabilities is essential for coordinated and accurate responses. Standardization enables the grouping of similar vulnerabilities, offering technical abstraction while clearly defining the underlying issue. This structure is particularly important for vulnerability scanners, as it allows them to report findings in a consistent and actionable manner.

A. Vulnerability Standardization

A global effort to standardize vulnerability denomination has been made, involving various initiatives, conventions, and established practices in this field. An example of this is the idea of a Common Weakness Enumeration (CWE), which is a uniform compilation of software weaknesses that are widespread across multiple systems and versions. Every CWE is assigned a distinct identification number, in the format CWE-(ID), allowing for easy identification and more fluid communication between cybersecurity personnel. The CWE website[1], upheld by MITRE, provides a compilation of CWEs containing weakness descriptions, examples, mitigations, and related CWEs.

When a vulnerability, related to one or more CWEs', is assigned to a particular version of software or service, the result is a Common Vulnerabilities and Exposures (CVE). Each CVE also has its own distinct identification number, in the format YYYY-(ID), where YYYY references the year of the discovery of the CVE and the ID is given in order of discovery. The CVE system offers a standard way to identify publicly recognized information-security vulnerabilities and exposures. MITRE runs a website[2] focused on CVEs, offering in-depth information on every vulnerability, such as descriptions, references, and potential impacts.

A Common Vulnerability Scoring System (CVSS) score is linked to each CVE to help prioritize and evaluate risks of different vulnerabilities. The rating is determined by criteria such as how easy it is to exploit, the level of complexity of the exploitation, the privileges needed by the attacker, the amount of user interaction required, and the possible impact on an organization. CVSS scores are frequently integrated with contextual and temporal metrics to derive a risk score tailored to a particular organization. A CVSS and impact can be calculated freely using calculators such as the one provided by NIST[3]

OWASP Top 10, a compilation created by OWASP[4], details the ten most severe security threats to companies. This list can serve as a reference for businesses looking to improve the security of their products and organizations against cyber threats. This list is updated every 3 to 4 years, depending on changes in the cybersecurity landscape, by various specialists reaching a consensus, and it is one of the most efficient ways to initiate a change in software development culture within an organization, leading to the production of better and more secure code.

B. Automated Security Testing

There are several methods used in automated vulnerability detection, each one essential, since each finds vulnerabilities that others do not [7]. Different strategies do not make others obsolete but complement each other, providing a wider coverage of vulnerabilities, especially when applied to different phases of an application lifecycle where their strengths are best utilized. Different approaches are more attuned to finding different vulnerabilities, and also differ in resource consumption as well as the need for access to information about the target application.

1) Static Application Security Testing (SAST): SAST is a testing methodology based on full access to the source code of the target, making it a type of white-box testing. In white-box testing, the internal structure, design, and implementation of the application are fully known and utilized during the testing process. This allows tools and professionals to thoroughly analyze the provided code, and identify potential security flaws[5]. However, the primary limitation of SAST tools stems from their lack of context about the application. Since these tools do not compile or execute code, focusing solely on coding patterns, identifying vulnerabilities that require a dynamic environment to be exploited becomes more challenging. This can lead to SAST tools flagging flaws in secure code, generating a high number of false positives, and failing to detect intricate vulnerabilities that span multiple components of the application. Configuration issues are also rarely identified by these tools, as such issues are generally not present in the source code itself. Although SAST tools identify vulnerabilities, confirming their validity remains challenging, as the flagged code is not executed during the analysis to verify its exploitability. Despite these challenges, SAST tools offer significant advantages over other solutions. Being a white-box approach, SAST is one of the fastest testing techniques and excels at detecting common but well-known vulnerabilities, such as Cross Site Scripting (XSS) and SQL Injection (SQLI). This makes it particularly useful for fast-paced environments such as CI/CD pipelines, resource-constrained setups, or in combination with other tools for organizations aiming for maximum coverage.

2) Dynamic Application Security Testing (DAST): DAST is a methodology used when access to source code cannot be granted or when SAST methods fail to detect vulnerabilities in more dynamic applications. It is a type of black-box testing, where the internal workings of the application are unknown, and only its external behavior is analyzed. Black-box testing focuses on inputs and outputs, treating the application as a "black box" whose internal logic and structure are not visible to the tester. DAST tools analyze applications by simulating actual attacks on a running instance of an application. These tools are often language-agnostic, meaning a single tool can analyze a multitude of targets regardless of their programming language. They are also very useful in the process of information gathering, providing much information about the framework versions and more about the target system. These tools are often designed to support plugins and additional payloads developed by third parties. DAST tools tend to have longer scanning times, consume a lot of network and computing resources, and need two machines, one running the

[1]https://cwe.mitre.org
[2]https://cve.mitre.org
[3]https://nvd.nist.gov/vuln-metrics/cvss/v3-calculator
[4]https://owasp.org
[5]https://owasp.org/wwwcommunity/Source_Code_Analysis_Tools

target and one the scanner. Therefore, they are not well suited for fast-paced pipelines with constant reiterations and changes, but they are ideal in a pre-production environment that mirrors production, allowing for catching vulnerabilities before the final deployment. DAST tools can use two approaches [8]: A passive one, which includes just noticing flaws visible to any user who interacts legitimately with the application. These security flaws are usually noticed in HTTP headers and responses, often caused by misconfigurations or unwanted leaked information. The active one is where an attack is simulated, and information regarding the system is extracted by means of fuzzing.

Fuzzing is a security testing technique that involves generating a large number of test cases using malformed data injection to discover unexpected behaviors such as bugs, crashes, and vulnerabilities.[6]. These often use combinations of static values, known to commonly cause problems (such as zero, negative numbers, and large numbers), as well as random data. New generation fuzzers use genetic algorithms to link injected data and impact on the target, where AFLplusplus is a reference [7]. In the case of web applications, which comprise popular systems, to ensure that the application has knowledge of every entry point possible and a basic comprehension of the mechanisms of the target, two approaches are often used: One being the use of a proxy server, allowing the tool to intercept all communications between the target and users, while the other is making use of **web-spiders**. Web-spiders automatically index a target when supplied with URLs to be used as a starting point as well as a blacklist. Then, a HTTP request is sent to each of the starting points and the response analyzed. If any more paths are discovered, the process of sending a request and analyzing the response for paths is continued until all paths and URLs are found. Although web-spiders can identify resources the proxy cannot, session management and HTML can become a hindrance. If the spider is unable to identify session expiration responses, it can become unable to work correctly, and HTML forms with strict validation can also hinder the enumeration. Also, in modern web applications, AJAX or GRAPHQL calls are used to modify the presentation layer, and spiders often fail in discovering complete paths.

3) Benchmarks: There are several approaches and vulnerable websites that can be used as benchmarks, with the most widely adopted being **OWASP Juice Shop**[8], **WAVSEP**[9] [9], **Damn Vulnerable Web Application**[10] [9]–[11] **OWASP WebGoat**[11] [10] and **OWASP Benchmark**[12] [12]. **OWASP Juice Shop** is a modern, intentionally vulnerable web application built with Node.js, Express, and Angular. It is widely used for security training, awareness demonstrations, CTFs, and as a test environment for security tools. **WebGoat** is another deliberately insecure application designed for developers to test vulnerabilities commonly found in Java-based applications that use widely adopted components. Although it is not yet fully optimized as a benchmark, the project team is actively working to expand its capabilities. **WAVSEP**is a deliberately vulnerable web application designed to evaluate the quality, feature set, and precision of web vulnerability scanners. It is particularly notable for including test cases that can lead scanners to generate false positives. **Damn Vulnerable Web Application** is a PHP/MySQL vulnerable web application to be used as a legal target for security professionals who want to test and improve their skills and toolkit. The **OWASP Benchmark** was selected because it is specifically designed to evaluate the accuracy, detection rate, and speed of automated vulnerability detection tools, rather than treating benchmarking as a secondary feature of a broader project. This dedicated focus makes it more effective in testing complex scenarios, including edge cases and common oversights, leading to a more precise and reliable assessment of each tool's capabilities. Additionally, it automates the evaluation process by parsing the output of security scanners, and directly comparing the detected vulnerabilities with the known existing ones. This eliminates the need for manual verification of found vulnerabilities or the development of custom parsing and verification scripts, ensuring consistency while significantly reducing effort. Furthermore, it generates detailed scorecards that provide clear, visual insights into each tool's performance. These features make OWASP Benchmark not only more precise in its evaluations but also better suited to fulfill its intended purpose as a comprehensive benchmarking tool.

III. Materials and Methods

Over the years, many strategies and visions regarding the ideal characteristics of a vulnerability scanner have led to the development of numerous open-source and licensed tools, each with distinct approaches, architectures, features, configurations, and outputs. Despite the availability of advanced solutions, many organizations (particularly those with limited budgets) struggle to justify the costs associated with vulnerability detection and management [6], [13]. As a result, open-source and free tools often emerge as the most viable option, as they reduce financial barriers and facilitate the adoption of defect management practices. Consequently, the tools selected for this study are open-source.

A. Scanning tools

In this section, an overview of the most widely used open source vulnerability scanners is presented. The tools were obtained from literature reviews, drawing on the work of S. Alazmi *et al.* [14] and D. Cruz*et al.* [3].

Arachni[13] is a DAST scanner based on Ruby, made to test web applications. This tool can alter its behavior based on the target. This allows for a pre-analysis of the results and, based

[6]https://owasp.org/www-community/Fuzzing
[7]https://aflplus.plus/
[8]https://demo.owasp-juice.shop/#/
[9]https://sectooladdict.blogspot.com
[10]https://github.com/digininja/DVWA
[11]https://owasp.org/www-project-webgoat/
[12]https://owasp.org/www-project-benchmark/
[13]https://github.com/Arachni/arachni

on a multitude of factors, predict the trustworthiness of the results provided. Arachni is able to scan client-side code as well as dynamic web applications, detecting the changes that occur during the crawling phase of the attack and adjusting itself to those. This allows for the discovery of patterns, assets, and attack vectors that might otherwise not be identifiable. Although still relevant, Arachni has not received any updates since last year, with SCNR[14] being its successor as a paid alternative.

W3af[15] is a modular vulnerability scanner with high coverage based on plugins written in Python, able to identify many different vulnerability types. W3af provides in-depth documentation that ranges from introductory topics to advanced use cases. It offers a graphical user interface and a wide range of plugins, with three different types: crawl plugins, which find new URLs, forms, and other injection points; audit plugins, which take advantage of the found injection points and send specific payloads to the server to identify vulnerabilities; and attack plugins, which can exploit the found vulnerabilities to gain access to unauthorized assets.

Skipfish[16] is a web application security reconnaissance tool developed with the help of security engineers at Google. Skipfish's purpose is to address common problems found in more used vulnerability scanners such as Nikto and Nessus. Advantages of using Skipfish include, but are not limited to, high performance, ease of use, handling of multi-framework websites, and content analysis. However, the lack of recent updates might pose some challenges. Security tools need to evolve to address new vulnerabilities and threats, making Skipfish, with its last update 12 years ago, a lackluster choice compared to other options.

Wapiti[17] is an open-source, free-to-use solution for black-box automated penetration testing. Wapiti offers the possibility of being used in automated tasks in scenarios of continuous testing. Although effective, Wapiti does not act like a MITM proxy, which means that it cannot find scripts where AJAX is involved. The author encourages switching to ZAP for more in-depth and effective penetration testing.

Vega[18] is a free and open-source web security scanner developed by Subgraph. Vega is written in Java and features a graphical user interface, making it accessible across multiple platforms including Linux, OS X, and Windows. Vega includes an automated scanner that can crawl websites, analyze page content, and find injection points. It also features an intercepting proxy that allows for tactical inspection and interaction with client-server communications, including SSL interception for HTTP websites. Vega enables users to extend their basic functionalities, integrating with new attack modules through an API in JavaScript.

Nikto[19] is an open-source web server scanner whose objective is slightly different from other solutions. Nikto is not only a security scanner but also an information-gathering tool. During its scans, it probes misconfigured files, permissions, and inputs, but also gathers information that might not be directly relevant but is useful for penetration testers. Nikto is not a stealthy tool by default, testing a web server in the quickest (and most resource-heavy) way, making its presence obvious in intrusion detection and prevention systems. However, there are some libraries that try to hide its presence in certain systems. Nikto runs in any environment containing Perl, as it is built on LibWhisker2.

Nuclei[20] is a fast vulnerability scanner that can scan not only applications but also cloud platforms and web infrastructure. Nuclei differs from other scanners by using YAML templates to mold the behavior of the scanner through the different phases of an attack. With its wide range of features and versatile use cases, Nuclei is an useful tool for professionals in every information technology sector.

GoLismero[21] is a Nikto-bundled vulnerability scanner that is platform-independent, running on Linux, Windows and OS X. This scanner unifies the results of well-known security tools into a readable, consistent output across all platforms. GoLismero uses commonly used standards to report the found vulnerabilities, providing the criticality of the vulnerabilities as well as a short description and possible solutions to mitigate the problem. It also encourages the community to develop its own plugins to extend the functionalities of the base application by using Python, a language with multiple convenient libraries and easy-to-learn.

OWASP ZAP[22] is one of the most popular and actively maintained web application security scanners. It is an open-source DAST tool developed by OWASP and designed for finding vulnerabilities in web applications. ZAP is suitable for beginners and experts alike, providing a user-friendly graphical interface alongside automation capabilities via scripts and APIs. Its features include intercepting proxy capabilities, automated and manual testing modes, and a wide array of plugins and extensions available through the ZAP Marketplace. ZAP supports complex scenarios such as handling authentication, scanning AJAX applications, and integrating into CI/CD pipelines, making it a robust choice for organizations aiming to implement comprehensive web application security testing.

B. Vulnerability assessment system

The primary factors considered by developers when selecting these solutions are **speed**, **precision**, **load** and **detection rate**. **Speed** refers to the time required for a website to be fully scanned. **Precision** measures how many of the detected vulnerabilities actually exist, while **detection rate** represents the number of existing vulnerabilities that the scanner is capable of identifying. **Load** refers to the impact that the scanning process has on the target system in terms of resource consumption, performance degradation, and potential service

[14]https://ecsypno.com/pages/codename-scnr
[15]https://github.com/andresriancho/w3af
[16]https://github.com/spinkham/skipfish
[17]https://wapiti-scanner.github.io
[18]https://subgraph.com/vega/
[19]https://github.com/sullo/nikto
[20]https://docs.projectdiscovery.io/tools/nuclei/overview
[21]https://github.com/golismero/golismero
[22]https://www.zaproxy.org

disruptions. Load is the amount of processing power, memory, bandwidth, or other system resources consumed during the scanning process. A scanner with a high load may generate excessive network traffic, CPU usage, or database queries, which can slow down or even temporarily disrupt the target application. These metrics are also not independent; for instance, a scanner with higher detection rate tends to have lower precision since increasing the range of detected vulnerabilities also raises the likelihood of false positives, where non-existent issues are incorrectly reported.

These aspects can be decisive depending on the development environment and the function the scanner fulfills. The **speed** of the scanning process is especially relevant for organizations integrating these tools into their CI/CD pipelines, as slower scanners can compromise the efficiency of the pipeline, and a comprehensive scan for each addition to the codebase is unnecessary and can be detrimental. **Precision** can be a deciding factor on less mature organizations where experts have more urgent tasks and cannot expend much time investigating vulnerabilities unless there is a high degree of confidence that they exist. Conversely, more mature organizations with well-established security teams and greater expertise may be more confident in handling a scanner with higher detection rate, as they possess the necessary resources to investigate every potential vulnerability effectively. **Load** is particularly crucial for organizations running resource-constrained systems or production environments where performance and availability must be maintained. High-load scanners can introduce latency, increase infrastructure costs, or even cause temporary outages, which may not be acceptable in certain business-critical applications. Finally, **detection rate** will be essential for organizations that implement this tool as a second or final layer of security, particularly when security is a critical feature of the product or system and it possesses the necessary resources to investigate every potential vulnerability. To achieve this, a benchmark will be conducted in which a deliberately vulnerable website will be scanned using all selected tools. The vulnerabilities identified by each tool will then be compared against the known existing vulnerabilities, allowing for the measurement of the three key characteristics.

To represent these four key characteristics, eight relevant metrics were measured: average and peak CPU usage, as well as average and peak memory usage, to assess system load; scan duration (HH:MM:SS), to evaluate time efficiency; **detection rate**, defined by the True Positive Rate or recall (TPR), calculated as:

$$TPR = \frac{TP}{TP + FN} \tag{1}$$

and **precision**, derived from the inverse of the FPR (False Positive Rate), calculated as:

$$FPR = \frac{FP}{FP + TN} \tag{2}$$

It is important to note that OWASP Benchmark calculates these rates per vulnerability category. As a result, scanners that detect a smaller number of vulnerabilities but cover a broader range of vulnerability types can achieve a higher TPR. This reflects their ability to address more categories even if the total count of detections is lower. Additionally, OWASP Benchmark provides a score to quantify a tools' effectiveness based on the ratio of true positives to false positives.

$$Score = TPR - FPR \tag{3}$$

The benchmarking process was performed using two different machines: one running the benchmark and the `sar` command[23] to record CPU and Memory usage; and the other running the scanner. The target machine is equipped with an Intel Xeon Gold 5218R CPU featuring 4 cores and 8GiB of RAM while the scanning machine has the same CPU and 16GiB of RAM. This process was then complemented using a shell script that timed execution times for each scanner.

For those requiring crawling, Hakrawler was executed first, and its discovered URLs were supplied as input. W3af and Skipfish, which initially faced dependency issues that made them unusable, they were instead run using a Docker image to ensure a stable and controlled environment. All scanners were run with their default configurations in active mode, to ensure consistency and avoid introducing variability based on manual tuning, as fine-tuning each tool falls outside the scope of this evaluation. While this approach allowed for a standardized comparison, it is important to acknowledge that the absence of custom configurations may have affected the performance of certain tools.

Some scanners offer extensive configuration options that could potentially improve precision, speed, or detection rates in real-world scenarios. However, since the goal of this evaluation is to assess the baseline capabilities of each tool, default settings provide a fair and reproducible comparison.

To visualize how different scanners perform under various conditions, multiple metrics were computed. These include distinct evaluations aimed at highlighting how tool performance can vary depending on the relative importance assigned to each characteristic.

To compare the performance of the different tools, we established the following formula that takes into account the load measured during the scan and the time it takes for it to execute.

$$PerformanceImpact = (Load - Baseline) \times Duration \tag{4}$$

In the tests the load was measured in % of the host resources used. For the impact calculation we used the averages of the relevant metrics during the scanner's execution. The time was converted to hours to apply the formula.

IV. Results and discussion

During the evaluation, several unexpected setbacks were observed. Arachni and Vega became largely inoperable, while

[23] https://www.man7.org/linux/man-pages/man1/sar.1.html

w3af, despite being somewhat functional due to third-party fixes, continued to face persistent dependency issues, making it unsuitable though it remains as a point of comparison. Additionally, Golismero and Nikto were mainly focused on detecting web-server vulnerabilities, rather than application vulnerabilities.

While these tools identified some vulnerabilities, those modules proved ineffective and failed to detect a single correct web-vulnerability. However, both can still be used alongside other scanners to improve detection rates, complementing scanners that specialize in detecting vulnerabilities in web applications. Another limitation was that some scanners, such as Nuclei and Nikto, lack built-in crawling capabilities, which restricted their ability to independently discover attack surfaces. To mitigate this, the open-source web crawler Hakrawler[24] was used to provide URLs to these scanners, ensuring they were not excluded due to this limitation.

The results reveal great disparities in performance and characteristics, significant trade-offs between speed, precision, and resource utilization were identified. For instance, while some tools prioritized rapid scanning at the expense of accuracy, others achieved broader vulnerability coverage but imposed substantial computational burdens. This suggests that each scanner may perform optimally in different situations, where requirements vary.

It's important to clarify that the number of false positives is not merely the count of irrelevant or unverified vulnerabilities, but specifically those triggered by test cases deliberately crafted to resemble real vulnerabilities without actually being exploitable. These deliberate traps assess a scanners' ability to differentiate real from false signals

A. Study results

The results presented in Table II lead to several key observations regarding the performance, efficiency, and detection capabilities of different scanners. ZAP and Wapiti emerged as the top-performing scanners, each excelling in different aspects of vulnerability detection. ZAP stands out due to its superior speed and higher detection rates, making it a robust choice for rapid and extensive scanning. In contrast, Wapiti demonstrates higher precision while consuming significantly fewer system resources, making it an efficient alternative for scenarios where accuracy and lower computational overhead are prioritized.

Despite their comparable overall scores, these two scanners do not identify the same vulnerabilities uniformly. Notably, Wapiti failed to detect any Insecure Cookies vulnerabilities, whereas ZAP achieved a detection rate of 64%. Conversely, in detecting Path Traversal vulnerabilities, Wapiti performed substantially better, achieving a detection rate of 54% compared to 11% obtained by ZAP. This divergence in detection capabilities suggests that using both these scanners together would be beneficial—not only to verify findings but also to increase overall vulnerability detection rates. Another important result is the efficiency of scanners regarding resource consumption. The two fastest scanners, Nuclei and Skipfish, displayed significant limitations in consistent vulnerability detection, both presenting negative scores. While Skipfish failed to identify vulnerabilities effectively, leading to a score of -100%, Nuclei, despite its speed, produced a high rate of false positives, resulting in a score close to zero. However, in fast-scanning situations where false positives are not a critical concern, Nuclei might still be a viable option.

[24]https://github.com/hakluke/hakrawler

TABLE I
SCANNER PERFORMANCE

Scanner	True Positives	False Positives	Precision	Detection Rate	Score
Wapiti	442	0	100.00	20.87	20.87
W3af	226	0	100.00	12.59	12.59
ZAP	403	6	99.76	22.15	21.91
Nuclei	22	19	99.26	0.74	0.00
Skipfish	0	1	0.00	0.00	−100.00

TABLE II
SCANNER IMPACT

Scanner	CPU (Avg/Peak)	Memory (Avg/Peak)	Time (h:m:s)	Performance Impact
Baseline	00.21 / 02.99	36.42 / 36.49	–	–
Wapiti	03.03 / 12.73	36.60 / 36.96	05:31:16	934
W3af	01.28 / 25.75	36.73 / 37.23	25:01:01	1606
ZAP	11.50 / 70.20	37.05 / 47.35	01:00:51	686
Nuclei	11.28 / 20.43	37.67 / 37.84	00:19:01	210
Skipfish	47.79 / 53.14	36.63 / 36.69	00:09:34	455
Nikto	17.25 / 17.96	37.63 / 37.65	00:00:10	2
Golismero	00.54 / 05.05	37.63 / 37.63	00:01:33	0

Additionally, the comparison reveals an interesting trend in resource utilization. While memory consumption remains relatively stable across different scanners, CPU utilization varies significantly depending on the tool used. For instance, ZAP exhibited the highest peak CPU usage at 70.2%, likely due to its extensive scanning capabilities, while lightweight tools like Golismero and Wapiti had considerably lower CPU peaks. These results indicate that organizations selecting a scanner must balance performance requirements with resource constraints, as more comprehensive tools tend to impose a higher computational load.

Finally, two scanners designed to assess server vulnerabilities rather than those present in web applications produced promising results. They demonstrated low resource consumption, and minimal impact on the target system, making them suitable for use alongside other scanners when a broader assessment is required. However, since these scanners are not compatible with the current benchmarking method, determining the best performer is not possible. Nikto is the most likely

candidate due to its frequent updates, ongoing maintenance, and active support.

In order to properly evaluate the scanners we attempt to establish a formula to compute a score that takes into account the OWASP Benchmark Score and complements it with the impact metric. By incorporating an exponential penalty

$$\text{Score\_Exp} = \text{Scanner Score} \times e^{-\alpha \times \left((\text{Load} - \text{Baseline}) \times \text{Duration}\right)},$$

we can ensure that scanners producing high load or long scan durations are penalized more steeply. This maintains a fair comparison between scanners that might have comparable detection metrics but very different performance footprints.

Alternatively, one can weight the denominator linearly

$$\text{Score\_Ratio} = \frac{\text{Scanner Score}}{1 + \beta \times \left((\text{Load} - \text{Baseline}) \times \text{Duration}\right)},$$

which provides a smoother, more intuitive adjustment. Here, a small overhead yields only a slight reduction in score, while large overheads grow the denominator quickly and reduce the overall rating. Both formulas aim to highlight scanners that effectively detect vulnerabilities without imposing excessive performance costs. In both formulas, α and β are both constants that adjust the weight of the performance impact on the final score calculation.

Figure 1 presents a visual representation of the various suggested metrics, using α and β values of 0.01. Since the OWASP Benchmark score can result in negative values, the final calculated scores were adjusted by replacing any negative values with 0 to improve readability.

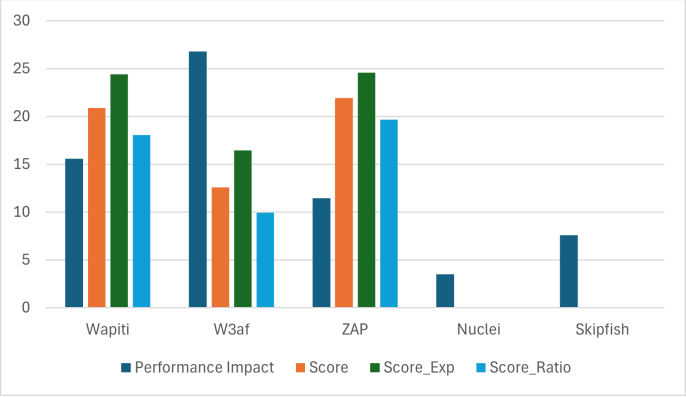


Fig. 1. Comparison of scanner performance.

B. Open challenges

While this study provides initial insights into the operational trade-offs of vulnerability scanners, further research remains essential to equip decision-makers with comprehensive selection criteria. Two research directions become critical to address current limitations:

Benchmark Enhancement: OWASP Benchmark focus on code-level vulnerabilities, creating gaps in assessing scanners' ability to detect infrastructure vulnerabilities. Future studies should integrate test cases for server specific vulnerabilities such as authentication flaws, insecure configurations and lack of updates—attack vectors frequently exploited in production environments. This expansion would allow for evaluations to mirror real-world attack targets, particularly for tools like Nikto that specialize in infrastructure scanning. In addition, scoring systems that allow for a better comparison such as F-score should be implemented.

Configuration Impact Analysis: The default settings evaluation, while revealing fundamental characteristics of each scanner, can overlook potential optimization through tool configuration. Focused experimentation with top performers (ZAP and Wapiti) could quantify how plugin integration, detection rule tuning, and scan policy modifications affect the observed speed-precision-resource triad. Systematic testing of ZAP's marketplace add-ons against Wapiti's extensible rule engine may identify configuration sweet spots where expanded detection capabilities maintain acceptable operational overhead — a critical consideration for enterprises requiring both thoroughness and infrastructure stability.

Expand Evaluation to Commercial Solutions: This study particularly investigates the use of freely available, open-source tools to increase transparency, configurability, and affordability—concerns that are of particular importance for organizations with limited financial means. Future research should incorporate commercial vulnerability scanners into the analysis. An in-depth comparison of commercial tools like Burp Suite Professional, Acunetix, and Tenable.io with open-source alternatives will clarify unsolved questions about the added value of proprietary software.. By grounding such discussions in empirical performance data, researchers will be able to provide fact-based guidance for hybrid toolchain development—namely, under which conditions open-source scanners can substitute proprietary tools in order to reduce costs.

V. Conclusions

The systematic evaluation of DAST tools shows operational trade-offs that influence their production environment viability. High-coverage scanners exemplified by OWASP ZAP achieve superior vulnerability detection rate (22.15% TPR) but have intense resource consumption, reaching 70.2% peak CPU utilization. Conversely, precision-focused tools like Wapiti have perfect accuracy at the expense of limited detection rate (20.87% TPR), requiring supplemental scanning for comprehensive coverage.

Practical implementation strategies emerge from the analysis: continuous integration pipelines benefit from rapid scanners like Nuclei despite higher false positive rates, provided secondary verification mechanisms exist. Resource-constrained environments favor efficient tools such as Wapiti (3.03% avg CPU), though periodic deep scans remain essential. High-security deployments require combined toolchains leveraging ZAP's breadth and Wapiti's precision to mitigate individual blind spots, particularly for cookie security and path traversal vulnerabilities.

Future research should investigate optimal configuration profiles for differing environments for the top-performing scanners (ZAP and Wapiti) through systematic testing with alternative detection libraries, plugin combinations, and scanning presets. A comparative study of ZAP's marketplace add-ons versus Wapiti's payload customization capabilities could establish whether configuration adjustments resolve observed trade-offs between detection rates and operational impact. Such experimentation would provide practical guidance for tuning scanners to specific organizational requirements while maintaining production system stability.

REFERENCES

[1] ISACA, "State of cybersecurity 2024," October 2024.

[2] N. I. of Standards and T. (NIST), "Technical guide to information security testing and assessment," National Institute of Standards and Technology (NIST), Tech. Rep., 2008.

[3] D. B. Cruz, J. R. Almeida, and J. L. Oliveira, "Open source solutions for vulnerability assessment: A comparative analysis," *IEEE Access*, vol. 11, pp. 100 234–100 255, 2023.

[4] M. R. Gajula and V. G. Vassilakis, "Evaluating the performance open-source vulnerability scanners," in *2024 14th International Symposium on Communication Systems, Networks and Digital Signal Processing (CSNDSP)*, 2024.

[5] E. Lavens, P. Philippaerts, and W. Joosen, "A quantitative assessment of the detection performance of web vulnerability scanners," in *ACM International Conference Proceeding Series*, 2022.

[6] K. Abdulghaffar, N. Elmrabit, and M. Yousefi, "Enhancing web application security through automated penetration testing with multiple vulnerability scanners," *Computers*, 2023.

[7] S. Elder, N. Zahan, R. Shu, M. Metro, V. Kozarev, T. Menzies, and L. Williams, "Do i really need all this work to find vulnerabilities?" *Empirical Software Engineering*, 2022.

[8] B. Rajić, Ž. Stanisavljević, and P. Vuletić, "Early web application attack detection using network traffic analysis," *International Journal of Information Security*, vol. 22, no. 1, pp. 77–91, 2023.

[9] B. Zukran and M. M. Siraj, "Performance comparison on sql injection and xss detection using open source vulnerability scanners," in *2021 International Conference on Data Science and Its Applications (ICoDSA)*, 2021.

[10] M. Y. Darus, M. Farhan Bin Bolhan, A. Kurniawan, Y. Muliono, C. R. Pardomuan, and M. Mohamad Hata, "Enhancing web application penetration testing with a static application security testing (sast) tool," in *2023 IEEE 8th International Conference on Recent Advances and Innovations in Engineering (ICRAIE)*, 2023.

[11] A. Al Anhar and Y. Suryanto, "Evaluation of web application vulnerability scanner for modern web application," in *2021 International Conference on Artificial Intelligence and Computer Science Technology (ICAICST)*, 2021.

[12] B. Mburano and W. Si, "Evaluation of web vulnerability scanners based on owasp benchmark," in *2018 26th International Conference on Systems Engineering (ICSEng)*, 2018.

[13] N. Alomar, P. Wijesekera, E. Qiu, and S. Egelman, "You've got your nice list of bugs, now what? vulnerability discovery and management processes in the wild," in *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*, 2020.

[14] S. Alazmi and D. C. De Leon, "A systematic literature review on the characteristics and effectiveness of web application vulnerability scanners," *IEEE Access*, 2022.

Red-Pi: An Adversary for the Water Sector

1st Agustín Javier Di Bartolo
Grupo de Ingeniería de Medios (GIM)
Universidad de Extremadura
Cáceres, Spain
adibartolo@unex.es

2nd Mohammadhossein Homaei
Grupo de Ingeniería de Medios (GIM)
Universidad de Extremadura
Cáceres, Spain
mhomaein@alumnos.unex.es

3rd Badreddine Abbaoui Toufiq
4rd Francisco Javier Muñoz Ruiz
5th Mar Ávila Vegas
Grupo de Ingeniería de Medios (GIM)
Universidad de Extremadura
Cáceres, Spain
{badreddine, franciscojmr, mmavila}
@unex.es

***Abstract*—In an increasingly digitalized world, companies in the water sector face challenges in protecting their technological assets. Many lack specialized IT or cybersecurity personnel and cannot afford periodic external audits, which are essential for identifying vulnerabilities and configuration errors. We present Red-Pi, a portable, accessible, and cost-effective solution designed to support water sector organizations in their digitalization process without requiring complex or expensive setups. Red-Pi connects to the internal network and performs automated penetration testing across industry-relevant protocols, generating daily reports with identified vulnerabilities, misconfigurations, and recommended mitigations. It operates autonomously and is remotely managed via a Telegram chatbot, enabling real-time updates and user interaction without technical expertise. In a real-world deployment, Red-Pi identified vulnerabilities in 20% of 60 devices within 3 hours, reducing the typical assessment time by over 60% and offering an estimated overall efficiency of 86%, highlighting its potential as a practical and scalable cybersecurity auditing tool for critical infrastructure.**

***Index Terms*—Cybersecurity, Water Sector, Red Teaming, Raspberry Pi**

I. Introduction

Water is critical to the operation of ecological, social, and economic systems. As with digital technologies that pervade our everyday lives, water cuts across policy areas and economic sectors. Europe's waters already face a multitude of challenges simultaneously, and these will be addressed by policy coherence and coordination that, although often called for, still remain elusive. Here, the water sector's digitalization can become a fundamental facilitator in balancing water policy and more efficient intervention. Meanwhile, citizens also request that the government service be operated most effectively and efficiently, with only justified and necessary spending. Thus, in a society based on water, digital transformation and digital service provision become significant elements in unifying the water sector into the overall paradigm of the digital economy [1], [2].

Digitalisation is a significant aspect of 21st-century governance that enables States to enhance the quality and scope of public services they deliver while, concurrently, enhancing the integrity of their operations for the good of society. Digitalisation is an enabler that fosters efficiency, flexibility, and transparency in the public sector, thereby helping to enhance the people's quality of life [3], [4].

There are several complexities and challenges that the phenomenon of digitalization portends. As urban systems and services continue to digitalize, there comes with it a rise in their exposure to inherent risks and vulnerabilities in cyberspace, thereby creating avenues for malicious actors and cyberattacks to penetrate essential sectors like Water and Sanitation (W&S), health, energy, and transport, among others [5].

In critical infrastructure, digitalized systems in the W&S sector are among the most sensitive due to their direct link to public health and hygiene controls. The sector is an important component of basic operations, such as the production and continuous supply of drinking water, as well as the collection, transmission, and treatment of wastewater—processes that are essential to maintaining sanitary conditions in both domestic and communal settings. Considering the life-critical nature of these activities, it comes as no surprise that a report released by the American Water Works Association has determined cyber threats to be the water and wastewater sector's most significant threat [6].

The water industry's legacy assets are often cited as a reason for not fully embracing digital transformation. These assets must be supported by new implementations, making it costly to rebuild systems from scratch. For example, when smart water meters are deployed on a large scale, IT systems require significant upgrades, sometimes leading to data silos. Legacy systems struggle to manage the changes brought about by digital transformation, as equipment typically has a lifespan of 7-20 years. Modern instrumentation can diagnose problems remotely, but cybersecurity issues currently prevent the full implementation of these systems.

The security of customers' data and implementing good cybersecurity practices are necessary in adopting digital solutions. The issue of cybersecurity is problematic in various industries, for instance, in the water industry, which handles sensitive customer data and is part of the national critical infrastructure. There are certain limitations, particularly on control systems that are not supposed to be internet-connected because of associated security risks. Customer data needs to

be secured because security breaches undermine trust and have economic ramifications. Data gathered by the water sector needs to be divided based on the risk involved. For instance, wastewater flow data does not have any appreciable security concerns; however, the moment this data affects control systems, it is a cybersecurity concern, a threat that is higher as digital transformation is more advanced [7].

This paper proposes the automation of an adversary for the W&S sectors using a Raspberry Pi, a powerful and portable device capable of performing reconnaissance, vulnerability analysis, and exploitation of various industry-oriented services and protocols. The system is remotely controlled with no human intervention, aiming to generate a daily security audit that produces a comprehensive report on vulnerabilities and potential solutions to improve the security of water sector companies.

The paper is structured as follows: Section II presents similar works to analyse the state-of-the-art, comparisons, limitations, and potential improvements. In Section III, we detail the methodology used, focusing on the design, architecture, and components required to build Red-Pi. Section IV presents the results obtained. Finally, Section V concludes the paper and discusses future work.

II. Related Work

A. Cybersecurity Challenges in the Water Sector

The water sector, as a critical component of national infrastructure, has increasingly become a target for cyber threats. The growing dependence on digital systems, including Industrial Control Systems (ICS), Supervisory Control and Data Acquisition (SCADA), Programmable Logic Controls (PLC), and Internet of Things (IoT) devices, has expanded the attack surface for cybercriminals. Unlike traditional IT networks, water infrastructure involves complex cyber-physical systems where cyberattacks can have severe consequences, such as disrupting water treatment processes, contaminating water supplies, or shutting down essential services.

To mitigate such risks, organizations have adopted red teaming exercises, where cybersecurity professionals simulate real-world cyberattacks to assess an organization's resilience. These exercises extend beyond conventional penetration testing by incorporating tactics such as social engineering, lateral movement, and persistent threats. The objective of red teaming in the water sector is to uncover vulnerabilities before real attackers exploit them, thereby strengthening defensive strategies and improving incident response capabilities.

Recent cyber incidents highlight the increasing risks facing water management companies. In many cases, attackers target confidential data belonging to organizations or their customers, aiming to extort payments or sell the stolen information on the black market. For instance, a British water utility company suffered the theft of 750 GB of sensitive data, affecting hundreds of thousands of customers and exposing corporate documents and personal records [8]. Similarly, in Texas, a water treatment company experienced a security breach that resulted in the loss of 33,000 files containing client information [9]. These incidents emphasize the urgent need for enhanced security measures in the water sector.

B. Low-Cost Cybersecurity Solutions via Pi

The growing need for cost-effective cybersecurity solutions has led to the adoption of Raspberry Pi in both defensive and offensive security applications, particularly for SMEs and critical infrastructure sectors like water management. On the defensive side, [10] demonstrates its use as an AI-driven cyber defense system with cloud-based filtering and real-time threat detection. For offensive security, projects such as [11] and [12] showcase their role as a stealthy penetration testing implant, providing remote access and persistence in compromised networks. The "Offensive IoT for Red Team Implants" series by [13], [14] extends these capabilities with hardware implants and LoRa-based attack execution. Additionally, [15] presents a portable cybersecurity toolkit for Wi-Fi auditing and penetration testing. These studies highlight the Raspberry Pi's affordability, versatility, and efficiency in cybersecurity operations.

C. Raspberry Pi in Water Security

Most research on Raspberry Pi-based cybersecurity focuses on general IT security, but its use in the water sector is not well studied. As cyber threats to water utilities increase, Raspberry Pi could be a low-cost solution to improve protection. A "Virtual Security Department" using Raspberry Pi can act as an automated red teaming system, helping SMEs assess risks continuously, especially in industries with limited cybersecurity staff and budgets. This approach provides a cost-effective way to secure critical infrastructure. Prior studies show that Raspberry Pi-based platforms can support both offensive (red teaming, penetration testing) and defensive (SIEM, IDS, anomaly detection) cybersecurity tasks. This is especially useful for SMEs and water sector operators who have limited resources but still need to protect ICS/SCADA systems from cyber threats.

1) Defensive Use: Low-Cost IDS/IPS: A study [10] shows how a Raspberry Pi can be used for an Intrusion Detection System (IDS) and Intrusion Prevention System (IPS) in IoT networks. The research uses open-source tools like Snort, Suricata, and Zeek to monitor traffic, detect anomalies, and reduce security risks. The study concludes that while Raspberry Pi-based IDS/IPS solutions are not as powerful as enterprise security devices, they offer a low-cost and scalable option for small networks and organizations with limited budgets.

2) Offensive Use: Red Teaming & Exploitation: Several studies and projects have shown how Raspberry Pi can be used for penetration testing and Red Team operations:

- Penetration Testing Dropbox with Raspberry Pi 4: A project by Artifice Security that turns Raspberry Pi 4 into a covert hacking device. It can be placed in a target network to collect information, find security weaknesses, and perform remote attacks.
- Raspberry Pi as a Penetration Testing Implant: Research by System Overlord explains how Raspberry Pi can be

TABLE I
COMPARISON OF RASPBERRY PI-BASED RED TEAMING AND CYBERSECURITY PROJECTS

Ref	Advantages	Goals
[10]	• Uses Raspberry Pi as an affordable security monitoring tool • Integrates open-source IDS/IPS solutions • Provides real-time network threat detection	• Develop a cost-effective IoT security framework • Improve intrusion detection capabilities in SMEs • Enable continuous network monitoring
[11]	• Portable penetration testing implant • Uses Raspberry Pi 4 for security assessments • Enables remote red teaming operations	• Deploy a versatile hacking device for penetration testing • Enable automated attack simulations • Enhance offensive security operations
[12]	• Stealthy security implant for red teaming • Remote control via secure communication channels • Supports automated security testing	• Develop a covert penetration testing tool • Improve adversarial emulation techniques • Enhance stealth capabilities for red teaming
[13]	• Uses Raspberry Pi Zero W for IoT exploitation • Implements hardware implants for red teaming • Explores new attack vectors in IoT environments	• Research covert red teaming techniques • Expand IoT hacking capabilities • Test offensive security tools in real-world scenarios
[14]	• Enhances red teaming tools with remote execution • Expands the operational range of Raspberry Pi implants • Provides over-the-air execution of attack scripts	• Improve prior hardware implant designs • Enable real-time attack execution via IoT devices • Expand attack simulation capabilities
[15]	• Fully integrated penetration testing suite • Wireless auditing and cybersecurity toolkit • Modular and portable offensive security tool	• Create a portable hacking suite for red teaming • Conduct real-world cybersecurity assessments • Utilize Raspberry Pi for network security testing

used as a hidden cyber attack tool. It runs Kali Linux, allows remote access, and uses network monitoring tools like Wireshark to capture data.

- Offensive IoT for Red Team Implants: Black Hills Information Security (BHIS) created a Raspberry Pi Zero W-based tool for Red Team hacking. They later improved it by adding LoRa communication, making remote attacks even more powerful.
- Ultimate Portable Hacking Suite with Raspberry Pi Zero W: A Raspberry Pi-based portable security tool designed for Wi-Fi hacking, penetration testing, and social engineering.

These projects show how Raspberry Pi is a cheap but powerful tool for cybersecurity testing and offensive security research.

3) Sector-Specific Adaptation: While most existing projects focus on generalized cybersecurity applications, our proposed approach introduces a specialized use case for the water industry. By leveraging Raspberry Pi-based security solutions, we aim to establish a "Virtual Security Department" capable of conducting continuous cybersecurity assessments in water management infrastructures.

This solution is particularly relevant for industries lacking dedicated security personnel and substantial cybersecurity budgets, as it provides:

- Automated red teaming operations to identify vulnerabilities in SCADA/ICS networks.
- Cost-effective IDS/IPS deployment to enhance threat detection in remote water facilities.
- Portable and scalable security monitoring solutions for critical infrastructure networks.

The integration of low-cost Raspberry Pi security appliances into water sector cybersecurity frameworks reinforces the practicality of automated, affordable cybersecurity solutions, particularly for SMEs and underfunded critical infrastructure entities [10], [11].

By synthesizing these prior works, it becomes clear that a Raspberry Pi-based platform can integrate both offensive (red teaming) and defensive (SIEM, IDS, anomaly detection) capabilities cost-effectively. This is particularly valuable for SMEs and water sector operators who are faced with constrained resources yet must address evolving cyber threats in ICS/SCADA environments (Table I).

III. METHODOLOGY

A. Design of the Red-Pi

This chapter explains the theoretical framework underlying Red-Pi, the self-sustaining opponent designed to defend stakeholders in the water sector. The following are some features that need consideration in its operation and design:

- Portability: The device must be portable, compact, and power-efficient. This will provide easy deployment to other sites without necessarily impacting resources or infrastructure significantly.
- Connectivity: The system must have the ability to establish numerous connections, including Ethernet, Wi-Fi, Bluetooth, and 5G. With this feature, Red-Pi can seamlessly fit into different network environments and cope with different operating contexts.
- Operating System: The device must have a Linux distribution, preferably Debian, Ubuntu, or Kali. These distributions have been noted for their stability, security features, and extensive support in the cybersecurity industry, thus being suitable for the tasks Red-Pi will be undertaking.
- Resources: The device must possess an adequate CPU and ample RAM to carry out the intended attack operations. The processing and memory must be enough to ensure efficient and quick operations.
- Storage capacity must be enough to accommodate the operating system and maintain reports completed on a monthly basis (30/31 days). It must be big enough to

make sure the performance of the device is not hampered and the data is maintained constantly available.
- Autonomy: Red-Pi must be able to perform its given tasks independently, in a rational order, without the intervention of an operator. This is to ensure smooth operation and minimum dependency on staff for supervision.
- Affordability and Accessibility: The chosen apparatus should be affordable and easily accessible in the operating area. This requirement is imperative to make Red-Pi an affordable and economically viable option for organizations operating in the water sector, thereby allowing for mass deployment if needed.
- Command and Control: It must be possible to monitor and command the device remotely, particularly if an inspection becomes a necessity because of a problem or for the purpose of software updating. Remote administration is facilitated by this feature; hence, there is no need for presence.
- Programming Languages: It must be able to support programming languages such as Bash and Python. Both of these languages find wide applications within the cybersecurity profession for executing reconnaissance, attack, and report-generation scripts so that Red-Pi can effectively execute them. Red-Pi's design responds to the demand for a low-cost, independent, and flexible solution that offers robust cybersecurity protection in the water industry, where security is paramount to the functioning of critical services.

B. Architecture and Components

Based on the outlined features, a comparison of three Raspberry Pi models is conducted to determine the most suitable option for the project. From Table I and Table VIII, the key conclusions are as follows:

- Raspberry Pi 3: It does not fulfill the basic requirements, as its RAM size, Ethernet port speed, USB versions, and operating systems supported are not acceptable for the project. It is thus eliminated as a possible option.
- Raspberry Pi 4: This model is compliant with the laid-down requirements and is highly appropriate to the project needs in the first version. It has ample processing capacity and accommodates the required operating systems, thus making it a sound choice for the development of Red-Pi.
- Raspberry Pi 5: Similar to the Raspberry Pi 4, this model also fulfills the project's requirements while having an extra benefit: It has the potential for future model upgrades if additional RAM is required or if a graphics processing unit would be necessary for more complex functionality, such as cracking hashes. This feature provides more room for flexibility and development.

Raspberry Pi Zero, Raspberry Pi Zero W, and Raspberry Pi Zero 2 W models have been excluded from consideration in this research because, although they fulfill some of the minimum parameters such as portability, power consumption, and low price, they are weak in various aspects, namely processing capacity, RAM capacity, and number of ports. These weaknesses necessitate their exclusion from the comparative research [16].

We need to acknowledge that we will be adding more elements to the chosen Raspberry Pi 5 to increase its connectivity options:

- 4G/5G USB SIM Adapter: This addition will deliver mobile network connectivity, complementing the Raspberry Pi's connectivity where other networks are unavailable or unreliable.
- 2.4/5 GHz Wireless USB Adapter (in passive mode) is designed to deliver enhanced Wi-Fi connectivity with increased flexibility in environments where high-performance wireless connections are needed.
- Low Range Bluetooth Adapter: This adapter enables Bluetooth connectivity, thereby enhancing the ease of communication with devices needing this standard, including sensors and peripheral devices.
- LORA Adapter plays an essential role in facilitating long-distance communications in situations where conventional networks may be restricted, thereby expanding the scope of connectivity alternatives.
- Zigbee Adapter: This module will allow the Raspberry Pi to be interfaced with devices using this specific communication protocol, especially in IoT applications where numerous sensors and devices must be connected via low-power networks.

C. Integration with Existing Infrastructure:

A key feature of Red-Pi is its ability to seamlessly integrate with the client's current infrastructure, which should be simply added without requiring the client to make any other configurations. For this reason, connectivity is a basic point, both for auditing the client network and for linking to the internet in order to report the network status and to send a daily report.

Red-Pi offers various means of connectivity. It can connect via Ethernet to start its network reconnaissance activities, and where Ethernet ports are not available, it can even connect via WIFI. A 5G SIM card adapter is utilized for the internet, which minimizes the risk of network blocks or depleting the client's internet. Bidirectional communication is done through a Telegram chatbot, in which users can track the status of each step and append reports upon completion. It also provides additional functionalities that are manageable through buttons in the chat. These operations enable the device status to be controlled and monitored independently without VPN or laptop connection, with the advantage of operating it in such a way that even the non-technical staff members within the company at Figure I.

D. Programming Language

For automation, Bash and Python were used together due to their simplicity, power, and our experience with them. This combination allows both easy scripting and handling of complex tasks. The approach follows a modular attack

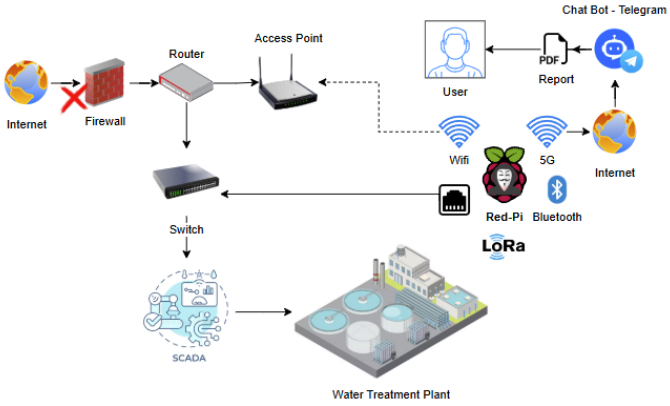


Fig. I. Architecture of Red-Pi

framework, where results from reconnaissance are reused, and attack modules target specific services and versions. The final output is a system report detailing findings, weaknesses, and possible fixes. This modular design makes updating Red-Pi easy. Once an attack or CVE is automated, the corresponding module is applied whenever that vulnerability is detected. The results are then logged and included in the final report.

E. Attack Automation

Before the beginning of the programming phase, flowcharts were created to outline the process in every reconnaissance, vulnerability analysis, exploitation, and report generation phase.

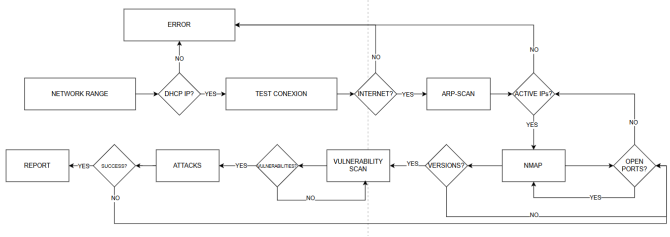


Fig. II. Flow Chart

As shown in Figure II, reconnaissance and attack were initially automated to allow the development of a process that would be reusable by other modules, thus incorporating feedback into the program. Testing is necessary when evaluating different versions, services, protocols, and scenarios to minimize false positives because each network is unique. The scripts were initially developed on a virtual machine with Kali Linux, where all the tools and libraries were already pre-installed.

Once the testing stage was done on the virtual machine, the scripts were transferred to the Raspberry Pi 5, which had been previously hardened to lock it down to protect it against physical attacks [17]

F. Targets

Since this piece of work is industrial water industry-focused, the aim is to address many protocols, services, and vulnerabilities common in the industrial and OT (Operational Technology) environment, which are MODBUS, MQTT, LoRaWAN, BLE, ZigBee, SNMP, and RTSP. Additionally, we shall also address protocols common in IT environments and talk to OT devices, which include SSH, FTP, TELNET, MySQL, Postgres, HTTP/HTTPS, NFS, and SMB.

We also want to identify routers, firewalls, gateways, or IoT devices that are using default settings, which may provide access to other devices, such as a LoRa gateway or an access point to another network.

Some devices and techniques still in use are more than 20 years old, so care of an unique nature is essential to ensure the attacks are not overly intrusive or loud to avoid undesirable service denial or network disruption, or disruption of specific devices.

G. Application Context

Red-Pi was tested in two environments:

The first was a controlled test setup with a LAN network, physical hardware, and a server running multiple virtual machines via Docker. These machines hosted vulnerable microservices and common configuration errors, such as default credentials and public shared files. This setup helped fine-tune performance, detect vulnerabilities, and ensure reliability.

The second was real-world networks from different organizations that allowed Red-Pi to run in their infrastructures, providing a more practical evaluation beyond the test environment.

H. Recognition

The reconnaissance process is the most vital stage of penetration testing. If we leave out or don't notice some device, some service, some version, there is are chance the probability of finding an initial point of entry would be significantly reduced.

Apart from being the most important, it is usually the most time-consuming step, with spending a lot of time waiting for tools to develop results, which is a dead time for a security analyzer.

To discover active devices and determine different protocols, we tested the tools as shown in Table II.

TABLE II
TOOLS FOR DISCOVERY AND PROTOCOL RECOGNITION

Tool	Function
Arp-Scan	Discovery of active devices
Nmap	Discovery of active devices
Massecan	Discovery of active devices
Whatweb	Web recognition
Cutycap	Web page captures
SMBClient	SMB service recognition
Showmont	NFS service recognition

The outcome of this reconnaissance will include IP addresses, open ports, services and their versions, and snapshots

if web services are found, along with other required services to perform vulnerability assessment based on seen protocols.

I. Vulnerability Analysis

Based on the results obtained in the previous step, the next step entails a bid to identify vulnerabilities linked to the services discovered.

To conduct the vulnerability analysis, we evaluated tools listed in Table III:

TABLE III
TOOLS FOR VULNERABILITY ANALYSIS

Tool	Function
Nmap	General vulnerability analysis
Nessus	General Vulnerability analysis
OpenVAS	General Vulnerability analysis
Burp Suite	Web Vulnerability analysis
JoomScan	Vulnerability analysis for Joomla
WPScan	Vulnerability analysis for Wordpress

From this stage, we will receive more information about potential vulnerabilities discovered, which must then be tested to verify the existence of such vulnerabilities.

J. Default user discovery

Many networks still have devices with default factory credentials that have not been changed. Identifying these devices is important because they are the first defense point and can be easily exploited in a network attack. For custom dictionary-based brute force attacks, the tools used are listed in Table IV.

TABLE IV
TOOLS FOR BRUTE FORCE ATTACKS WITH CUSTOM DICTIONARIES

Tool	Function
Hydra	Brute force for ssh, ftp, telnet, mysql, and postgres
Medusa	Brute force for ssh, ftp, telnet, mysql, and postgres
WPScan	Brute force for Wordpress
Docker	Brute force for RSTP
Bash script	Brute force SNMP
Python script	Brute force MQTT
Crunch	Create custom dictionaries
Documentation	Wiki and online documents
SecList	Dictionaries Repositories

Small dictionaries are used in these attacks to avoid generating too much noise in the network. The goal is to test a list of usernames and passwords that match the detected device. For example, if a brand X device is found, its specific default credentials will be used. If the brand is unknown, a general dictionary will be used until Red-Pi includes a

K. Explotation

With all the collected information, we now attempt to exploit the misconfigurations and vulnerabilities found during reconnaissance and scanning. This stage is important because it tests if the discovered vulnerabilities can be used for unauthorized access or system damage. The tools used apply different attack methods, such as brute force and injection, to check for possible exploitation. Some of these tools are listed in Table V.

The results will confirm whether the vulnerabilities are truly exploitable, ensuring that only real threats are reported. This helps avoid false positives. By testing the exploits, we can assess the severity of the vulnerabilities and suggest solutions. This process not only evaluates system security but also helps improve the overall network defense strategy.

TABLE V
TOOLS FOR EXPLOTATION

Tool	Function
Responder	Obtend NTML Hash
Ettercap	Main in the Middle
MQTT Python Library	Read/Write MQTT
PyModbus Python Library	Read/Write Modbus
DB Python Library	Dump DB
Github Repositories	CVE Exploits
Metaesploit	Software to act like an attacker
Proxychain	Pivoting

The results obtained will validate whether the vulnerabilities are exploitable or not, helping to avoid false positives in the reports.

L. Report Generator

One of the key aspects of Red-Pi compared to other similar writings is the fact that it is capable of producing a report that records activities of what has been done on the reconnaissance procedure, vulnerability scanning, and the exploitation.

Once its daily routine comes to a close, Red-Pi creates a report in the following specifications:

Number of active devices that have been detected, including IP, MAC, and Vendor. For each IP address, it produces a report containing the following details:

- Services, ports, and versions were identified.
- If the web service is running, a screenshot is included.
- Vulnerabilities discovered.
- Potential proposals or solutions to reduce the determined vulnerabilities.

Finally, the report is sent automatically through the Telegram API to access the acquired results quickly and easily.

IV. RESULTS

To assess the timing and performance, an evaluation of Red-Pi was conducted on a production network consisting of approximately 60 distinct devices, with no prior knowledge of the network being tested.

In Table VI below, the activities performed and the corresponding time taken by Red-Pi are listed.

The times it will take are proportional to the devices in the network, and also the services it identifies; differences can therefore be felt between different configurations. Less service-based networks or fewer devices may take a simpler process to work on, whereas larger or more complex networks may require more time to complete. The efficiency and swiftness of Red-Pi are hence significantly influenced by network complexity.

As can be seen in Figure III, Red-Pi can notify users via Telegram about the execution of its tasks and the time taken

TABLE VI
TASK AND TIME RESULTS

Task	Times (sec)
TEST_CONN	0,05
ARP_SCAN	2,05
NMAP_TCP	6735,97
NMAP_UDP	20,60
NMAP_SV	818,20
SERVICE_SEPARATOR	0,03
WEB_SCREEN	249,08
WEB_RECON	51,40
BRUTE_FORCE_SSH	153,36
BRUTE_FORCE_FTP	60,40
BRUTE_FORCE_MYSQL	120,02
BRUTE_FORCE_POSTGRE	80,10
MODBUS_SNIFF	420,42
PLC_ATTACK	354,12
SCADA_ATTACK	1064,05
MQTT_SNIFF	10,71
RESPONDER_HASH	880,60
REPORT	110,60
TOTAL	10.699,57

for execution. Using this feature, users can receive real-time feedback about the ongoing security audit being conducted, which eliminates manual checkups. This notification keeps the user informed in real time about the status and allows them to respond if needed.

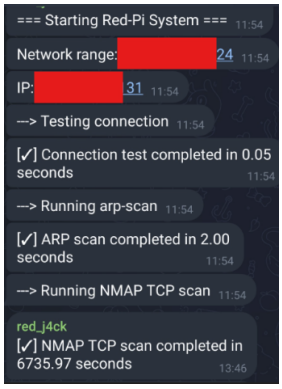


Fig. III. Newtwork Recognition

Table VII details the devices and vulnerabilities detected in the test network.

TABLE VII
DEVICES AND VULNERABILITIES DETECTED

Device Type	Brand	Service	Vulnerability	Quantity
Gateway IoT	Draguio	WEB	Default credentials	2
Access Point	TP-Link	WEB	Default credentials	1
Device IoT	Raspberry	SSH	Default credentials	2
PLC	OpenPLC	WEB	CVE-2021-31630 (RCE)	1
SCADA	SCADABR	WEB	CVE-2021-26828 (RCE)	1
Device IoT	Raspberry	MQTT	Unprotected MQTT	1
Router	Mikrotik	WINBOX	Weak credentials	1
Server	SuperMicro	SSH/WEB	Default credentials	4
PC	Windows	NTML	Hash extraction	1
PC	Windows	SMB	Unprotected shared folders	2
Server	Linux	NFS	Unprotected shared disk	1

The network certainly had a large number of exposed devices and misconfigurations, which exposed open services that made them vulnerable.

After completing the comprehensive list of attacks, Red-Pi formats the report, appends it, and sends it over Telegram, as shown in Figures IV and V. The automated process of report generation and sending is time-saving when compared to manual reporting and effectively sends the results of the security audit timely manner, providing vital insights into the security posture of the network.

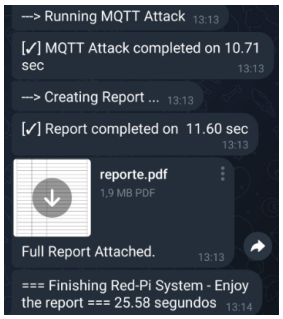
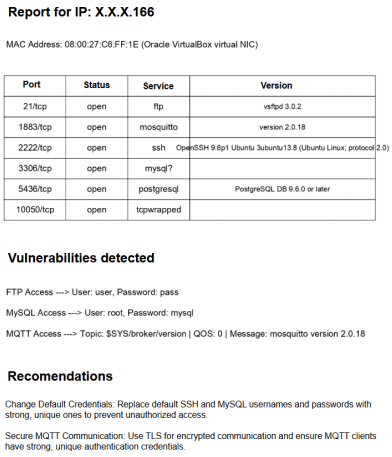


Fig. IV. Full Report Attached for Red-Pi


Report for IP: X.X.X.166

MAC Address: 08:00:27:C8:FF:1E (Oracle VirtualBox virtual NIC)

Port	Status	Service	Version
21/tcp	open	ftp	vsftpd 3.0.2
1883/tcp	open	mosquitto	version 2.0.18
2222/tcp	open	ssh	OpenSSH 9.6p1 Ubuntu 3ubuntu13.8 (Ubuntu Linux; protocol 2.0)
3306/tcp	open	mysql?	
5436/tcp	open	postgresql	PostgreSQL DB 9.6.0 or later
10050/tcp	open	tcpwrapped	

Vulnerabilities detected

FTP Access ---> User: user, Password: pass

MySQL Access ---> User: root, Password: mysql

MQTT Access ---> Topic: $SYS/broker/version | QOS: 0 | Message: mosquitto version 2.0.18

Recomendations

Change Default Credentials: Replace default SSH and MySQL usernames and passwords with strong, unique ones to prevent unauthorized access.

Secure MQTT Communication: Use TLS for encrypted communication and ensure MQTT clients have strong, unique authentication credentials.

Fig. V. Full IP Report

A. *Efficiency Evaluation of Red-Pi*

Red-Pi's overall effectiveness is measured using a Composite Efficiency Score (CES) that integrates four weighted factors: detection accuracy, execution time, cost efficiency, and automation level. The formula is:

$$\text{CES} = w_1 E_{\text{detection}} + w_2 E_{\text{time}} + w_3 E_{\text{cost}} + w_4 E_{\text{automation}} \quad (1)$$

With the values: $E_{\text{detection}} = 0.20$, $E_{\text{time}} = 0.70$, $E_{\text{cost}} = 0.97$, $E_{\text{automation}} = 1.00$, and weights $w_1 = 0.2$, $w_2 = 0.3$, $w_3 = 0.3$, $w_4 = 0.2$, we compute:

$$\text{CES} = 0.2(0.20) + 0.3(0.70) + 0.3(0.97) + 0.2(1.00) = 0.861 \quad (2)$$

This results in an estimated 86.1% efficiency, demonstrating Red-Pi's value as a fast, low-cost, and fully automated cybersecurity auditing tool.

TABLE VIII
COMPARISON OF RASPBERRY PI 3, 4, AND 5 SPECIFICATIONS

Feature	Raspberry Pi 3	Raspberry Pi 4	Raspberry Pi 5
CPU	Quad-core Cortex-A53 @ 1.2 GHz	Quad-core Cortex-A72 @ 1.5 GHz	Broadcom BCM2712, Quad-core Cortex-A76 @ 2.4 GHz
GPU	VideoCore IV	VideoCore VI	VideoCore VII (Supports OpenGL ES 3.1 and Vulkan 1.2)
RAM	1 GB LPDDR2	2 GB, 4 GB, 8 GB LPDDR4	4 GB, 8 GB LPDDR4X
Wi-Fi Connectivity	802.11n (Wi-Fi 4)	802.11ac (Wi-Fi 5)	Wi-Fi 5 (802.11ac)
Bluetooth	4.2	5.0	5.0 / BLE (Bluetooth Low Energy)
USB Ports	4 × USB 2.0	2 × USB 3.0, 2 × USB 2.0	2 × USB 3.0, 2 × USB 2.0
Ethernet Port	10/100 Mbps	Gigabit Ethernet	Gigabit Ethernet with PoE optional
Storage	MicroSD	MicroSD	MicroSD slot, option for SSD M.2 (via optional HAT)
PCIe and Expansion	Not available	Not available	1 × PCIe 2.0 x1 (hardware expansion)
Compatible OS	Raspberry Pi OS, Ubuntu, Windows IoT Core	Raspberry Pi OS, Ubuntu, Fedora, Kali Linux	Raspberry Pi OS, Ubuntu, Fedora, Kali Linux
Other Features	Not available	Not available	Power button, RTC (Real Time Clock), optional accessories
Power Supply	5V/2.5A	5V/3A	5V/3A
Power Consumption	Low (5W)	Moderate (7W)	Moderate (8W)
Approximate Price	$60	$90 (4 GB)	$99 (4 GB) - $140 (8 GB)

V. CONCLUSION

Red-Pi has demonstrated its effectiveness in automating security testing across various industrial and IT protocols, identifying vulnerabilities and generating reports without requiring cybersecurity expertise. Its efficiency increases significantly in complex infrastructures. In a real-world test involving 60 devices, Red-Pi identified vulnerabilities in approximately 20% of them in under 3 hours, greatly reducing the time and effort compared to manual assessments. Red-Pi minimizes the exposure of vulnerable systems and reduces dependence on specialized personnel. Its portable design and integration with Telegram for real-time reporting make it especially useful for small organizations with limited security resources. Planned improvements include expanding attack modules, reducing scan time, minimizing detection noise, and integrating DeepSeek for context-aware recommendations. Additional support for communication protocols such as LoRa, BLE, and Zigbee will enhance Red-Pi's applicability in broader industrial and IoT scenarios, reinforcing its value as a practical, low-cost cybersecurity solution for critical infrastructure.

ACKNOWLEDGEMENT

This initiative is carried out within the framework of the funds of the Recovery, Transformation and Resilience Plan, financed by the European Union (Next Generation) - National Cybersecurity Institute (INCIBE) in the project C107/23.

REFERENCES

[1] R. E. Agency, "Digitalisation in the water sector recommendations for policy developments at eu level," 2022, accessed: 2025-02-25. [Online]. Available: https://op.europa.eu/en/publication-detail/-/publication/6eb837b2-54df-11ed-92ed-01aa75ed71a1

[2] M. Homaei, O. Mogollón-Gutiérrez, J. C. Sancho, M. Ávila, and A. Caro, "A review of digital twins and their application in cybersecurity based on artificial intelligence," *Artificial Intelligence Review*, vol. 57, no. 8, Jul. 2024. [Online]. Available: http://dx.doi.org/10.1007/s10462-024-10805-3

[3] Y. Laubshtein, *Protecting Water and Sanitation Infrastructure from Cyberthreats: A Cybersecurity Study for Latin America and the Caribbean*. Inter-American Development Bank, May 2023. [Online]. Available: https://doi.org/10.18235/0004876

[4] M. Homaei, A. J. Di Bartolo, M. Ávila, O. Mogollón-Gutiérrez, and A. Caro, "Digital transformation in the water distribution system based on the digital twins concept," 2024, available on arXiv. [Online]. Available: https://arxiv.org/abs/2412.06694

[5] G. Féry, "The digital journey of water and sanitation utilities in latin america and the caribbean: What is at stake and how to begin," Nov. 2022. [Online]. Available: http://dx.doi.org/10.18235/0004562

[6] J. Germano, "Cybersecurity risk & responsibility in the water sector," 2019, accessed: 2025-02-25. [Online]. Available: https://www.awwa.org/Portals/0/AWWA/Government/AWWACybersecurityRiskandResponsibility.pdf

[7] I. I. W. Association), "Global trends & challenges in water science, research and management," 2022, accessed: 2025-02-25. [Online]. Available: https://iwa-network.org/publications/global-trends-and-challenges-in-water-science-research-and-management/

[8] S. Water. (2025) Cyber attack update for customers. Accessed: 25-Feb-2025. [Online]. Available: https://www.southernwater.co.uk/latest-news/cyber-attack-update-for-customers/

[9] WISDIAM, "Recent cyber attacks on water and wastewater systems," 2025, accessed: 25-Feb-2025. [Online]. Available: https://wisdiam.com/publications/recent-cyber-attacks-water-wastewater/

[10] G. Palmer and J. Scott, "Enhancing iot security affordably with raspberry pi and open-source ids/ips," *International Journal of Cybersecurity and Digital Forensics*, vol. 12, no. 4, pp. 100–112, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10596202

[11] A. Security. (2023) Penetration testing dropbox with raspberry pi 4. [Online]. Available: https://artificesecurity.com/how-to-build-your-own-penetration-testing-dropbox-using-a-raspberry-pi-4/

[12] S. Overlord. (2023) Raspberry pi as a penetration testing implant (dropbox). [Online]. Available: https://www.systemoverlord.com/raspberrypi-dropbox

[13] B. H. I. S. (BHIS). (2022) Offensive iot for red team implants - part 1. [Online]. Available: https://www.blackhillsinfosec.com/offensive-iot-for-red-team-implants-part-1/

[14] ——. (2023) Offensive iot for red team implants - part 3. [Online]. Available: https://www.blackhillsinfosec.com/offensive-iot-for-red-team-implants-part-3/

[15] J. Miller. (2022) Building the ultimate portable hacking suite with a raspberry pi zero w. https://assume-breach.medium.com/building-the-ultimate-portable-hacking-suite-with-a-raspberry-pi-zero-w-dbc60704d872.

[16] R. P. Foundation, "Raspberry pi documentation," 2025, accessed: 2025-02-21. [Online]. Available: https://www.raspberrypi.com/documentation/computers/raspberry-pi.html

[17] M. C. Ghanem, E. Almeida Palmieri, W. Sowinski-Mydlarz, D. Dunsin, and S. Al-Sudani, "Weaponized iot: A comprehensive comparative forensic analysis of hacker raspberry pi and pc kali linux machine," feb 2025, preprint available at https://doi.org/10.20944/preprints202501.0203.v3.

Addressing the evolving threat of social engineering

Telmo Nicolas Sauce*, Luís Batista*, João Paulo Barraca†, and João Rafael Almeida*
* IEETA / DETI, LASI, University of Aveiro, Portugal
* IT / DETI, University of Aveiro, Portugal

***Abstract*—As digital infrastructures become increasingly integral to modern life, organizations face a growing array of cybersecurity threats that extend beyond traditional technical vulnerabilities. Among these, social engineering attacks stand out for their exploitation of human psychology, making them particularly challenging to detect and prevent. These attacks can bypass security measures and evade traditional defenses, making them particularly challenging to mitigate. To address these threats, institutions must not only enhance technical defenses, but also invest in comprehensive staff training, prevention strategies, and ongoing readiness assessments. However, keeping these programs updated to the ever-evolving social engineering landscape, along with conducting regular tests such as physical social engineering exercises or phishing campaigns, can be costly, time-consuming and disruptive to the normal working flow of the company. This paper examines the rising prevalence and sophistication of social engineering, highlighting its evolution into a major cybersecurity concern. We explore the stages of a social engineering attack, from initial research to post-exploitation, comparing established models such as Mitnick's cycle with more nuanced frameworks that account for evolving tactics.**

***Index Terms*—Cybersecurity, Social Engineering, Security Awareness, Threat Mitigation**

I. Introduction

As the world becomes increasingly digital, technology plays a crucial role in our daily lives, shaping how we work, communicate, and manage information. By digitalizing their operations, organizations leverage computers to optimize workflows, safeguard essential data, and remotely manage critical systems. Adopting these advancements directly affects systems' isolation, which would otherwise be internally managed, thus presenting reduced interaction possibilities [1]. Failures and disruptions on these systems can have devastating consequences, resulting from unintentional malfunctions or malicious intent. This second cause can lead to aggravated repercussions depending on the objective of the attacker, which, as technology evolves, so do their available attack vectors. This evolution of threats pushes traditional defense mechanisms to their limits, requiring institutions to adopt comprehensive and adaptive strategies to protect their assets against a wider range of risks [2].

A rising trend in cybersecurity constitutes social engineering attacks, which exploit human psychology rather than relying solely on technical vulnerabilities. For instance, a report by Positive Technologies indicated that social engineering accounted for 43% of incidents recorded in the first half of 2023, increasing to 57% in the first half of 2024 [3]. Similarly, the European Union Agency for Cybersecurity (ENISA) 2024 threat landscape report highlighted that social engineering remains a key tactic employed by cybercriminals across various sectors. Between July 2023 and June 2024, 28% of the observed social engineering incidents targeted the general public, followed by digital infrastructure (15%), public administration (10%), and the financial sector (10%) [4].

Threat actors can employ various techniques to deceive their victims into revealing sensitive information or gaining unauthorized access. Among the most popular digital social engineering tactics stands phishing. In this attack, adversaries send messages masquerading as trusted entities to lure end-users to spoofed/fraudulent websites that try to compromise their data. In addition to digital methods, physical social engineering attacks are becoming a significant threat. These bypass conventional cybersecurity measures to ultimately target the weakest link in the security chain, the human factor [5], [6].

Social engineering tactics can be combined with technical methods to create hybrid attacks, which are challenging to defend against. These allow cybercriminals to bypass early security measures and infiltrate organizations, which is done by leveraging social engineering to obtain initial access, followed by technical strategies to gain further control or exfiltrate data. Once successful, these hybrid attacks become hard to detect and mitigate, as they appear to originate from within the system itself. Although detection technologies have improved, they are far from flawless and often fail to keep pace with evolving attack methods. Attackers are constantly refining their tactics, which hinders detection and mitigation processes in automated systems. For institutions, this means that securing their digital systems is no longer enough. They must train their staff to recognize these threats, develop robust prevention processes, and continuously test their capacity to respond to these rapidly evolving attack vectors.

This paper presents an in-depth analysis of the evolving threat landscape, focusing on social engineering attacks and their integration with technical exploits. It explores how attackers take advantage of human behavior to bypass security mechanisms, allied with privilege escalation techniques used to accelerate system compromise.

II. Social Engineering Overview

Social engineering refers to all techniques that persuade individuals to reveal specific information or perform actions that compromise their data. It takes advantage of human limitations and their flawed nature, which is difficult to counter with standard hardware or software-based solutions [7], [8]. While awareness training can help reduce risks, people remain susceptible to manipulation, misplacing their trust, and

making it difficult for complete prevention [8]. The number of social engineering attacks has significantly increased in recent years, reaching 57% of the reported incidents in 2024, 11% more compared to 2023 [1]. In 2023, the Computer Emergency Response Team for Portugal (CERT.PT) reported that phishing/smishing incidents were the most frequently recorded, totaling about 35% of all cases. An additional 10% of occurrences were related to other types of social engineering incidents, with vishing prevailing with 35% of the subtotal, followed by CEO fraud with 31% [2].

A. Commonly Identified Manipulation Stages

Social engineering attacks come in many forms, following distinct patterns that unfold in several stages [8]. A widely recognized model that illustrates these phases is the **Social Engineering Attack Cycle** by Kevin Mitnick [9], represented in Figure 1. This model has four stages: **Research**, **Developing Rapport and Trust**, **Exploiting Trust**, and **Utilising Information**.

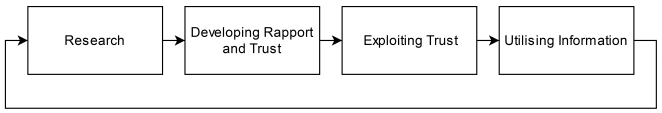


Fig. 1. Kevin Mitnick's Social Engineering Attack Cycle scheme

On the **Research** stage, the attacker tries to collect as much information as possible about the target. This is done to strengthen the attack vector for the following stages, as the quality of the obtained data increases the likelihood of establishing a strong and trusting relationship with the individual [10].

The next stage is the **Developing Rapport and Trust** of the victim, during which the attacker tries to create a strong relationship with the target. This stage is essential to increase the probability that the target will share the requested information [9]. Some strategies to achieve this trustworthiness include misrepresenting a legitimate identity, citing credible individuals associated with the entity, showing a need for assistance, or faking an authoritative role [10].

Next comes the **Exploiting Trust** phase, where the trusting relationship previously established between the malicious actor and the target individual is abused. In this, attackers usually try to emotionally influence the victim to provide sensitive data or make security mistakes [8], [10]. This can be accomplished by directly asking for information, requesting the target to perform certain privileged actions, or manipulating the user into seeking help [9].

Finally, **Utilising Information** is the phase where the attacker attempts to exit without leaving any trace [8]. Here, the previously gathered data is used to reach the attack's goal or to advance to further steps of the plan [10].

The authors of the **Social Engineering attack framework** [10] propose several modifications to the Kevin Mitnick model, as shown in Figure 2. They introduce an initial phase called **Attack Formulation**, during which the social engineer establishes a clear conceptualization of the attack and identifies a suitable target. Following the **Research** stage, the authors add a **Preparation** phase that focuses on consolidating the gathered data, crafting the attack vectors, and planning the attack. Next, they discuss whether the **Utilising Information** stage should be included in the social engineering attack. This reflection is related to the fact that the data has already been captured, and its utilization will constitute another type of attack. Therefore, they replace this phase with the **Debriefing** stage, at which the attacker should make the target return to the normal emotional state that was constructed during the **Exploitation Phase**. This step is essential to ensure that the target does not reflect on what happened, similarly to the **Impact** phase at the cyberattack lifecycle, and where the attacker tries to conceal their actions. The final step of the **Debriefing** phase is the **Transition**, during which the social engineer determines if the goal was achieved, or if it should return to the **Research** phase.

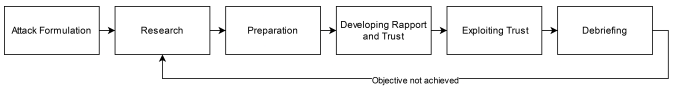


Fig. 2. Social Engineering Attack Framework scheme

B. Categorization of Attack Strategies

Social engineering attacks can be categorized in various ways depending on the perspective of the social engineer. Figure 3 gives a hierarchical overview of these various classifications.

1) Entity-Based Social Engineering: One way to categorize these attacks is by the type of entity that executes them, which can be subdivided into two categories: human-based or software-based. In **human-based** attacks, the attacker engages in direct communication to extract the desired information, typically done in person or by phone calls impersonations [8]. This attack relies heavily on the attacker's ability to manipulate the target through non-verbal cues, such as tone, appearance, and body language, making it challenging to scale or automate [8], [11]. While the scope may seem limited compared to software-based attacks, human-based attacks can adapt dynamically to different situations. Attackers can respond to subtle behavioral signals, such as hesitation or discomfort, and adjust their strategies in real time to build trust and reduce suspicion. **Software-based** attacks are carried out with devices or systems to automate the process, making it possible to target many victims simultaneously with minimal effort. These techniques allow attacks to reach thousands of targets within seconds, dramatically increasing the likelihood of success, since deceiving a single victim can produce valuable results [8]. Such attacks are highly efficient because they are automated and scalable, allowing attackers to achieve their goals without needing direct, real-time engagement.

[1]https://global.ptsecurity.com/analytics/cyberthreats-in-the-public-sector
[2]https://www.cncs.gov.pt/docs/rel-riscosconflitos2024-obcibercncs.pdf

2) Approach-Based Social Engineering: Attacks can also be classified by approach and divided into three categories: physical-based, technical-based, and social-based. **Physical-based** attacks involve the attacker performing physical actions to retrieve sensitive data from the target. Often used techniques include dumpster diving, where attackers search through dumpsters looking for valuable documents or information, such as passwords, and tailgating, in which the social engineer follows an authorized individual into a restricted area without proper identification or clearance [8], [12]. Such attacks can be advantageous because they completely evade digital security systems by targeting weak points in physical security. Additionally, physical access enables attackers to install malicious devices, such as rogue routers. These blend into the local network and can remain undetected for extended periods. **Technical-based** approaches are mainly carried out over the internet, with threat actors often using Open-source intelligence (OSINT) to gather information from the target's social media profiles and other public online resources [12]. One key advantage of this approach is that it is usually free, and the victim remains unaware of the malicious actor's actions, allowing them to gather valuable data without raising suspicion. **Social-based** attacks rely on creating relationships with the victim through interpersonal skills and psychological strategies to extract information [8]. Unlike technical-based or physical-based attacks, which may use tools or physical access, social-based attacks are centered on human emotions, including greed and curiosity. These emotions are frequently abused in tactics like phishing and baiting, where attackers create scenarios that deceive the target into sharing sensitive information or performing malicious actions against the organization.

3) Communication-Based Social Engineering: The final classification method focuses on how the social engineer communicates with the victim, divided into two categories: direct and indirect communication. **Direct communication** refers to attacks where the victim interacts with the social engineer through physical presence, eye contact, or voice. Techniques which utilize this type of communication are shoulder surfing, phone-based social engineering, and impersonation [8]. Direct communication can be further subdivided into bidirectional and unidirectional communication. **Bidirectional communication** occurs when both parties participate in the conversation. For instance, the attacker may impersonate an IT support employee and try to obtain credentials. In **unidirectional communication**, the hacker initiates the interaction while the victim passively receives the message without responding, for example, a voicemail without requiring feedback. Finally, **indirect communication** occurs when there is no communication between the attacker and the victim. Instead, third-party methods are used to manipulate the targets, which can be done through phishing emails or infected flash drives left strategically to lure victims into taking insecure actions, for example [10].

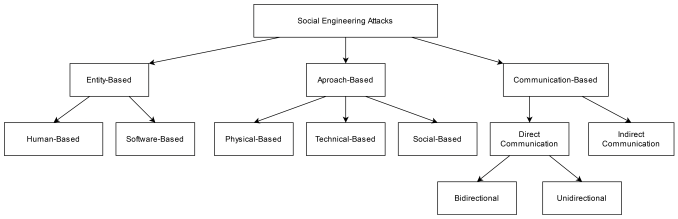


Fig. 3. Social Engineering Classification Tree

TABLE I
MAPPING OF ATTACK TECHNIQUES TO COMMUNICATION CHANNELS

Technique	Email	Website	Social Media	Cloud Services	IMA	SMS	Telephone VoIP	Physical
Dumpster Diving	-	-	-	-	-	-	-	X
Shoulder Surfing	-	-	-	-	-	-	-	X
Tailgating and Piggybacking	-	-	-	-	-	-	-	X
KeyStroke Injection	-	-	-	-	-	-	-	X
Keylogger	X	-	-	-	-	-	-	X
Evil Twin	-	-	-	-	-	-	-	X
Deauthenticator	-	-	-	-	-	-	-	X
Phishing	X	X	X	X	-	-	-	-
Spear Phishing	X	X	X	-	-	-	-	-
Whaling	X	-	X	-	-	-	-	-
Vishing	-	-	-	-	-	-	X	-
Smishing	-	-	-	-	X	X	-	-
Search Engine Phishing	-	X	-	-	-	-	-	-
Watering Hole Attack	-	X	-	-	-	-	-	-
Fake Software	X	X	X	X	X	-	-	-
Pop-Up Windows	X	X	-	X	-	-	-	-
DNS Spoofing	-	X	-	-	-	-	-	-
Impersonation	X	-	X	-	X	X	X	X
Pretexting	X	-	X	-	X	X	X	-
Quid Pro Quo	X	X	X	-	X	X	X	-
Reverse Social Engineering	X	X	-	-	X	X	X	-
Baiting	-	-	-	-	-	-	-	X

III. SOCIAL ENGINEERING TECHNIQUES AND ATTACK CHANNELS

Social engineers must carefully select their channels and techniques by considering factors such as the attack's objectives, the stage of the interaction, and the information already gathered about the target. By understanding the strengths and weaknesses of each channel, social engineers can enhance their chances of achieving a successful outcome.

In the context of social engineering, a channel refers to the medium or method used by the attacker to manipulate or deceive their target. Different channels can facilitate different outcomes, as some are better suited for gathering sensitive data, while others may be more effective for delivering malware or exploiting the target's trust. Some channels used to carry out these attacks include email, websites, social media, cloud services, Instant Messaging Applications (IMA), Short Message Service (SMS), telephone/Voice over Internet Protocol (VoIP), and physical means.

Table I provides an overview of the potential channels through which a malicious actor can exploit a victim [11], [12]. Each technique outlined is described in more detail in the next subsections, structured according to the approach-based classification discussed in II-B2.

A. Physical-Based Attacks

- **Dumpster Diving** refers to the act of searching through physical trash or discarded materials to uncover confidential or sensitive data that can be exploited for malicious intent [11]. Hackers typically look for documents that may contain personal data, proprietary information, or other sensitive materials.

- **Shoulder Surfing** involves an attacker observing a target to gather private information. This technique requires positioning behind the target to discreetly view their screen or keyboard, allowing the attacker to capture passwords or other confidential data without the victim's perception [12].
- **Tailgating and Piggybacking** are methods of gaining unauthorized access to secure areas. This can be achieved by following someone into a restricted location or taking advantage of their assistance, such as asking them to hold the door while claiming to have forgotten an access badge [8].
- **KeyStroke Injection** is a USB device, often disguised as a thumb drive, that automatically inserts keystrokes into any host computer it is connected to. This attack can be used to create a remote access to the victim's machine or even install malware on it.
- **Keylogger** is a malicious software or USB device that silently records every keystroke on a compromised machine. It records sensitive information entered by users that is transmitted to the attacker, such as passwords.
- **Evil Twin** is an unauthorized Access Point (AP) that simulates a given network. In this attack, the social engineer creates a Wi-Fi network with the same Service Set Identifier (SSID) of another known Wi-Fi network. This is done to lure victims into connecting to the fake AP and potentially jeopardize user credentials to the legitimate network. By keeping the user connected to the fake network, threat actors can further leverage Man-in-the-Middle (MITM) attacks to retrieve even more sensitive data from the user.
- **Deauthenticator** is a technique that is used to force users to disconnect from their current network. This is done by sending specific deauthentication frames to the devices, and is usually used in conjunction with rogue AP attacks to try to force the users to connect to the fake network and intercept sensitive information.

B. Technical-Based Attacks

- **Phishing** is a fraudulent technique in which attackers deceive individuals into divulging sensitive information or credentials through fake emails or messages that appear to come from trustworthy organizations, such as banks or reputable online services. These attacks usually aim to reach a broad audience to maximize their effectiveness [12], and have several variants:
  - **Spear Phishing** is a targeted form of phishing that focuses on specific individuals or organizations. These attacks leverage personal details to craft a convincing narrative, which significantly enhances their likelihood of success [13].
  - **Whaling** is a specialized form of spear phishing aimed at high-ranking officials or key decision-makers within an organization [13].
  - **Vishing and Smishing** are forms of phishing that rely on different communication methods. Vishing uses phone calls to manipulate victims into providing personal information, while smishing delivers phishing messages via SMS containing harmful links or a request for sensitive data [13].
  - **Search Engine Phishing** involves strategically placing fraudulent websites at the top of search engine results to deceive users. They may unintentionally enter personal information or download malware from these sites, which look like legitimate services [14].
- **Watering Hole Attack** targets specific groups by compromising the websites that they frequently visit. Once these are infected with malware, attackers can subsequently exploit users who open them [12].
- **Fake Software** attacks consist of creating fake login pages that closely resemble authentic websites. After users unknowingly provide their sensitive information, attackers can redirect them to the legitimate page, leaving victims unaware of the breach [8].
- **Pop-Up Windows** trick users into entering their login credentials through misleading alerts about connection problems or malware. These pop-ups might also present false solutions that activate malware instead of resolving the reported issue [8].
- **DNS Spoofing** involves altering DNS requests to redirect users from legitimate websites to fraudulent ones. This tactic typically employs cloned sites that capture sensitive information [14].

C. Social-Based Attacks

- **Impersonation** attacks involve an attacker disguising themselves as someone known to the victim, such as a colleague or authority figure. It exploits the victim's sense of familiarity and trust to share sensitive information or perform actions they otherwise would not. Frequently, it is used as a foundation for other methods, such as those described below.
- **Pretexting** consists of planning a fabricated story to persuade someone to share personal details. For example, an attacker might pretend to be an IT staff member requesting credentials to access critical systems [8].
- **Quid Pro Quo** operates on the concept of "something for something", where attackers promise attractive offers, such as special deals or giveaways, in exchange for personal data like usernames or email addresses [14].
- **Reverse Social Engineering** involves drawing the victim into making contact with the attacker. This approach makes victims more likely to share personal information because they initiated the contact themselves, reducing suspicion about the legitimacy of the interaction [11].
- **Baiting** takes advantage of curiosity or desire by leaving infected USB drives in public spaces or offering captivating free software for a limited time. When victims interact with these bait items, they may unknowingly install harmful software [14].

IV. Prevention and Mitigation Measures

Awareness programs are a vital aspect of an organization's strategy against social engineering, as they reduce the success rate of the attacks. These initiatives are designed to provide employees with the knowledge and skills necessary to identify, prevent, and respond to such threats. They should educate employees about the various types of social engineering attacks and train them to recognize suspicious behaviors, including unusual requests, unsolicited communications, or unauthorized attempts to access restricted areas. Furthermore, awareness programs should point out best practices in cybersecurity, including the use of strong passwords, proper disposal of documents, and cautious online behaviour, as these are fundamental to safeguarding organizational assets. To improve the outcomes, organizations must adopt interactive and engaging training methods that reflect real-world scenarios, such as simulated phishing campaigns and penetration testing. These approaches not only provide valuable insights into employee preparedness but also highlight areas that need improvement in the defense mechanisms and awareness programs [7], [8].

Phishing campaigns replicate attacks by sending malicious emails to employees, used to assess whether individuals correctly identify them as malicious or preemptively shared sensitive information with the security team. In this campaign, several metrics help the organization evaluate staff practices and the effectiveness of previous training. The **click rate** refers to the percentage of individuals who interact with the email, whether by clicking on a malicious link or downloading an attachment. The number of people who **report and inquire** about emails helps to measure the perceptiveness of staff to recognize and report suspicious threats. Ultimately, this enables cybersecurity teams to identify and mitigate threats more quickly. In case of a real threat, **reporting even after falling** can also offer key insights. If an employee has successfully been tricked into clicking a link, they must report it immediately to reduce the attack damage as much as possible. **Monitoring the data that is disclosed** helps the cybersecurity team understand what data employees are more susceptible to providing, such as personal information or sensitive company information like business plans, financial data, and others. **Tracking statistics over time** helps measure how well awareness training has improved employee vigilance. It provides insights on the most recognized types of attack, as well as which programs might not be yielding positive results [15].

By tracking these metrics, using tools such as GoPhish, the organization can determine if the results from the tests are positive, meaning that individuals are effectively identifying and avoiding manipulation attempts. In such a scenario, organizations can gradually increase the difficulty of these tests, ensuring that teams remain consistently challenged and continue to improve their ability to recognize attacks. On the other hand, a high click rate or interaction with phishing emails can pinpoint which departments are more susceptible to attacks, meaning they may require additional training. It also helps identify which types of phishing campaigns are more effective in each department, allowing the creation of targeted training programs to address specific vulnerabilities and improve overall security awareness.

Penetration testing is a controlled security assessment that simulates real attack scenarios to uncover vulnerabilities in technical or physical defenses, policies, and even employees' practices. By mimicking the tactics of a potential hacker, institutions gain valuable insight to address existing weaknesses before they can be exploited [16]. Hardware tools like the Wi-Fi Pineapple can further enhance these tests by demonstrating how unauthorized devices can infiltrate a system. Organizations can use these tools in controlled environments to evaluate employee responses, applying similar metrics as phishing campaigns.

It is crucial that companies do not stop conducting these tests, given the constantly evolving nature of cyber threats and social engineering techniques. Periodic assessments are essential to reinforce key concepts and skills by regularly reviewing test results. Institutions can adapt their training programs to address the weaknesses identified in each round of tests, ensuring that training and awareness programs remain effective.

To mitigate social engineering threats, organizations should implement technical controls such as phishing detection systems, firewalls, and monitoring software to reduce the likelihood of malicious content reaching employees. Additionally, implementing biometric and Multi-Factor Authentication (MFA) can help protect sensitive systems, even if an attacker gains access to user credentials. Finally, regularly reviewing and updating security measures is crucial to ensure that organizational defenses remain updated and effective against evolving threats [7], [8].

A. Challenges in Awareness Training Methods

Current awareness training methods typically rely on a **one-size-fits-all** approach, failing to account for variations in experience levels and individual susceptibility. As training must be provided to all staff, from high-level executives to janitors, its implementation becomes challenging due to the diverse technical backgrounds of the employees [17].

Social engineering tactics are continuously evolving as cybercriminals adapt to the preventive measures taken by institutions. As they develop new and more sophisticated techniques, staff become increasingly unfamiliar with these threats and struggle to recognize them. This issue is further aggravated by the lack of resources dedicated to cybersecurity, which means training programs generally fail to **keep up** with emerging threats. Moreover, organizations must regularly test employee readiness, but the difficulty in allocating staff to conduct training and develop modern programs, combined with the high costs of training, makes it challenging. Individuals often lack the motivation to participate in regular training, leading to a careless attitude that undermines the effectiveness of these programs. This ultimately leaves them more vulnerable to attacks and unaware of emerging threats [17].

One challenge that can leave organizations vulnerable to social engineering attacks is the tendency to focus on digital cybersecurity while underestimating the risks posed by physical social engineering. Organizations often allocate most of their security budgets to technological defenses and basic physical security measures, such as surveillance systems, guards, and locks. This imbalance, motivated by limited resources, usually leads to prioritizing the most common attack types and thus neglecting human factor susceptibilities.

The cost associated with physical penetration testing for these tests can be especially high for small institutions with limited budgets or for large organizations with multiple facilities to evaluate. Given the complexity of the tests, they often rely on external teams, which further adds to the financial concern. In addition to the cost, physical penetration testing can be disruptive and may cause unintended negative consequences, such as crashing outdated systems during the testing process. To minimize potential damage, testers should anticipate hazards and take preventive measures, such as shutting down critical operations. However, this can interfere with ongoing projects and processes, compromising their availability [18].

B. Low-Cost Penetration Testing Hardware Solutions

To address the challenges of high costs and logistical barriers in physical penetration testing, this work demonstrates the feasibility of building low-cost hardware tools to simulate physical social engineering attacks. These tools enable organizations to evaluate their resilience against threats like rogue Wi-Fi networks, malicious USB devices, and RFID cloning, using widely available components such as Raspberry Pi microcontrollers, ESP32 modules, and open-source software. These tools not only reduce reliance on expensive commercial solutions, such as Hak5, but also promote proactive security testing, particularly for small-to-medium enterprises and educational institutions with limited cybersecurity budgets.

1) Keystroke: A keystroke injection attack occurs when a victim unknowingly connects a malicious USB device to their computer. This device, often powered by a microcontroller or small single-board computer, such as an Arduino, Teensy, or Raspberry Pi, emulates a legitimate keyboard or mouse and sends unauthorized keystrokes or mouse movements to the victim's system. The attack exploits the USB Human Interface Device (HID) standard, which allows devices to communicate with a computer without requiring special drivers, as these are standardized across all major operating systems. Since operating systems inherently trust HID devices, a malicious USB can execute arbitrary commands on the target machine.

When a USB device is connected, the host system automatically queries it for information in a process called enumeration. The device provides descriptors that define its identity, specify power and interface settings, and establish the structure of input and output of data. If the device provides valid descriptor data, the host system accepts it and initiates communication, enabling data exchange through HID reports. These are structured data packets containing key presses or mouse movements, and their format is predefined during the enumeration phase. Once the host system recognizes the USB as an HID device, it begins accepting traffic from it without further verification. At this point, the malicious device can start executing commands as if they were sent by the user [19].

A reverse shell payload can be executed using keystroke injection as a gateway to compromise a victim's machine. This payload grants the attacker full control over the target system while requiring only a small set of keystrokes, making the attack faster and harder to detect. To carry out the attack, an HID device is used to send the keystrokes, while a server hosts the reverse shell connection for the victim's machine to connect back to. When the USB device is plugged in, the payload is executed, establishing a backdoor to the victim's system.

A reverse shell attack is a hacking technique where an attacker gains remote control over a compromised machine. Instead of the attacker directly connecting to the victim, the victim initiates the connection back to the attacker's machine, creating an interactive shell. This is a powerful attack because normal inbound connections are often blocked by firewalls, and Network Address Translation (NAT) hides private IP addresses, making devices unreachable from the internet. A reverse shell bypasses these challenges by leveraging the trusted outbound connections of the victim's enterprise network, rather than an unknown source attempting to gain unauthorized access.

The HID device can be implemented using the open source project pico-ducky [3]. This requires a Raspberry Pi Pico, a compact microcontroller without internet connectivity, and a micro USB to USB-A cable or dongle. Additionally, for development purposes, a breadboard and jumper wires are needed to disable the auto-run feature by connecting GP0 to a GND pin.

Pico-Ducky operates based on two core files: code.py and duckyinpython.py. By default, the microcontroller executes code.py, which starts by checking whether GP0 is connected to a GND pin to determine if it is in setup mode. If GP0 is not grounded, the script dynamically selects the payload based on which pin is grounded: GP4, GP5, GP10, or GP11, executing the payload.dd, payload2.dd, payload3.dd, or payload4.dd, respectively. Once the appropriate task is selected, duckyinpython.py interprets and executes the Duckyscript instructions. Alternatively, for this project, a custom payload was developed using Adafruit HID libraries [4] to execute a Python script. In this variation, only code.py runs at boot, providing additional functionality, including mouse movement support, an ability not present in the current version of duckyinpython.py.

The zero-hid project can deploy a keystroke injection attack, requiring a Raspberry Pi Zero and a micro USB-to-USB-A adapter or dongle. Unlike the Raspberry Pi Pico, which is a microcontroller, the Raspberry Pi Zero is a full-fledged computer with an operating system, enabling it to perform more complex tasks. The zero-hid tool utilizes Python for

[3]https://github.com/dbisu/pico-ducky

[4]https://github.com/adafruit/Adafruit_CircuitPython_HID

attack deployment but requires prior configuration to enable its HID functionality. The Raspberry Pi Zero must be set up as a USB gadget by installing the `dwc2` driver, which allows it to switch between host mode, acting as a computer, and gadget mode, behaving as a peripheral. Additionally, the `libcomposite` framework must be installed to enable flexible USB gadget configurations, allowing the device to function as multiple peripherals simultaneously. Furthermore, the `init_usb_gadget` binary is downloaded and registered as a systemd service to run on boot, ensuring that the Raspberry Pi Zero automatically loads the required drivers, creates the HID devices, and operates in gadget mode upon startup.

To execute a reverse shell attack on the victim's machine, a server is required. This must have a public IP address or be accessible from the victim's network. To simplify the deployment of the reverse shell attack, the reverse_ssh tool uses Docker to automate the process and generate URLs that can be executed on other machines. These URLs inject a malicious binary that establishes a reverse shell connection back to the server. A key feature of this project is its ability to maintain persistent connections with victims until either the victim terminates the connection or the administrator removes the client from the server.

To initiate the keystroke injection attack, the process begins by accessing the server via SSH and running the Docker container for the reverse_ssh project. Once active, the command `link` generates a URL, which is later used to deploy the attack on a target machine, typically a Windows system, though Linux is also an option. With the HID device set up, the generated URL is embedded in the payload file to properly reference the server. Since keystroke mappings differ across keyboard layouts, the appropriate configuration must be selected. Considering the targeted environment, the Portuguese layout was used.

Once the tools are properly configured, the next step is to either leave the device in a strategic location or insert it directly into a target machine. When plugged in, the device executes the payload within seconds, establishing a connection between the victim's machine and the reverse shell server. At this point, further access to the compromised machine could be gained.

2) Evil Twin: An Evil Twin attack occurs when an attacker introduces an unauthorized router or AP into a network. By routing traffic through the rogue router, the attacker can capture sensitive information, modify data in transit, or redirect users to malicious websites. This rogue device can be a small single-board computer like a Raspberry Pi or a microcontroller like the ESP32 or ESP8266. The attack exploits the users' trust in the network infrastructure, such as public or company-provided networks.

These APs can be named after trusted entities, such as "Library" or "Free-WiFi-Google" to appear legitimate. Once a user connects, the network may either provide internet access or remain offline. Regardless of connectivity, the system will track the total number of unique devices that connect, demonstrating the potential success of DNS spoofing, traffic sniffing, and other malicious activities without actively performing them.

The attack will first be implemented using an ESP32 device, a microcontroller developed by Espressif Systems, which features built-in Wi-Fi and Bluetooth capabilities. The ESP32 has excellent support through Arduino libraries, making it ideal for deploying APs, identifying nearby networks, and gathering information about connected devices, such as their MAC addresses. However, it cannot forward internet traffic to other networks due to the lack of built-in routing or NAT support. As a result, the ESP32 will create an AP, but it will not provide internet access.

In addition to the ESP32, a Raspberry Pi 5 can also be used in the attack. The Raspberry Pi is a more powerful and fully functional computer capable of acting as a router, unlike the ESP32. With the help of an external Wi-Fi dongle, it gains a second WLAN interface, allowing it to function both as a station and an AP. As a result, the rogue AP created by the Raspberry Pi will have internet access, as it can forward traffic.

To set up the network on the Raspberry Pi, hostapd, a Linux-based software for managing wireless networks, is configured to broadcast a malicious SSID, such as "Free-McDonalds-Wifi". The hostapd.conf configuration file is modified to define WPA2 security settings, further enhancing the network's authenticity. Once the rogue AP is established, dnsmasq is used to provide DNS and DHCP services, enabling devices to connect and access the internet. Finally, NAT rules are configured to allow all incoming and outgoing connections, making the rogue network appear like a legitimate internet service provider.

Alongside the rogue APs, a device connection logging script is implemented to track which devices connect to the network. This script utilizes the ARP-scan tool to monitor network activity by executing the arp -a command. This command retrieves the ARP cache, which contains a list of active devices on the local network, mapping their IP addresses to their corresponding MAC addresses. Each time a new device connects to the rogue network, the script detects its presence by identifying its MAC address. This information is then recorded and stored in a JSON file, allowing for further analysis of the number of devices that unknowingly connect.

The ESP32 deploys a script that activates the AP and logs the connected users. The device will then be strategically placed in locations where free Wi-Fi is common, such as campus borders, high-traffic user zones, and areas accessible to privileged personnel, which increases the potential impact of the attack. As users join the rogue network, their connections will be logged, and the attacker can retrieve these logs either by directly accessing the device or by connecting to the ESP32's network.

For the Raspberry Pi, the process begins with an SSH connection to the device, linking it to the Eduroam network, and setting up its AP. Forwarding capabilities are then configured to facilitate network interactions. Like the ESP32, the Pi 5 should be placed in busy locations to maximize its effectiveness.

3) RFID Attack: A Radio-Frequency Identification (RFID) sniffing attack intercepts wireless communication between RFID cards and readers by capturing the radio signals emitted during transactions. This technique allows attackers to stealthily extract sensitive institutional card information, potentially enabling them to bypass physical security measures. Once stolen, data can be cloned onto a blank RFID card or emulated using a programmable device, allowing unauthorized access to secured areas or fraudulent transactions. This attack can be performed on any device with NFC capabilities, including microcontrollers such as the RC522 or PN532.

RFID systems often grant access when a card transmits a valid Unique Identifier (UID) or stored credential data to a reader, making them vulnerable to sniffing attacks due to weak security mechanisms. Many common RFID cards, such as MIFARE Classic and Ultralight [20], transmit this data unencrypted, allowing any NFC-capable device to passively capture the UID by emulating a legitimate reader. For example, an attacker could place a counterfeit reader near the legitimate one, causing both devices to receive the card's signal simultaneously. Since these systems typically verify access by checking only the UID against a whitelist, without cryptographic authentication or challenge-response mechanisms, an attacker can intercept the plaintext transmission, clone it onto a blank card, and gain unauthorized access. Moreover, the system cannot distinguish the cloned credential from a legitimate one.

An RC522 RFID reader module can be used to read and write data on RFID cards, in combination with an Arduino ESP32 Nano, to facilitate communication. The Arduino can be programmed using the RFID library [5], which implements SPI protocols to interact with the RC522 module, handling tasks such as card authentication, UID extraction, and sector-level data reads/writes. After connecting the RC522 module to the Arduino using a breadboard and jumper wires, a program was deployed to read and store the UID of RFID cards in the Arduino's flash memory.

To carry out the attack, the malicious RFID device is placed near a legitimate RFID reader. When a card is scanned, both the legitimate and malicious readers capture its UID. While the legitimate reader processes the request, either granting or denying access, the malicious device stores the UID. The attacker then retrieves the malicious RFID gadget, extracts the captured UID, and programs it onto a rewritable RFID card, effectively cloning a legitimate access card. Finally, the cloned card can be used to access restricted areas corresponding to the privileges of the original card.

V. Conclusions

Social engineering remains one of the most significant threats to organizational security as it targets the human element. Strategies, such as implementing tailored awareness programs, conducting simulated phishing campaigns, and using penetration testing, demonstrate the importance of aligning organizational defenses with real-world attack scenarios.

[5]https://github.com/miguelbalboa/rfid

Organizations must continuously adapt their strategies to address emerging threats, integrating both digital and physical security measures to mitigate risks. Developing affordable tools for physical penetration testing and updating training programs to reflect the latest social engineering trends are critical for maintaining an adaptive and resilient security posture.

References

[1] E. Knapp and J. Langill, *Industrial Cyber Security History and Trends*, 12 2015, pp. 41–57.

[2] T. Victor-Mgbachi, "Navigating cybersecurity beyond compliance: understanding your threat landscape and vulnerabilities," *Iconic Research and Engineering Journals*, vol. 7, 2024.

[3] P. Technologies, "Cyberthreats in the Public Sector," 2024, accessed: 2025-04-10. [Online]. Available: https://global.ptsecurity.com/analytics/cyberthreats-in-the-public-sector

[4] European Union Agency for Cybersecurity (ENISA), "Enisa threat landscape 2024," European Union Agency for Cybersecurity, Tech. Rep., 2024.

[5] K. L. Chiew, K. S. C. Yong, and C. L. Tan, "A survey of phishing attacks: Their types, vectors and technical approaches," *Expert Systems with Applications*, vol. 106, pp. 1–20, 2018.

[6] Z. Alkhalil, C. Hewage, L. Nawaf, and I. Khan, "Phishing attacks: A recent comprehensive study and a new anatomy," *Frontiers in Computer Science*, vol. 3, p. 563060, 2021.

[7] P. P. Parthy and G. Rajendran, "Identification and prevention of social engineering attacks on an enterprise," in *2019 International Carnahan Conference on Security Technology (ICCST)*. IEEE, 2019.

[8] F. Salahdine and N. Kaabouch, "Social engineering attacks: A survey," *Future internet*, vol. 11, no. 4, 2019.

[9] K. D. Mitnick and W. L. Simon, *The art of deception: Controlling the human element of security*. John Wiley & Sons, 2003.

[10] F. Mouton, M. M. Malan, L. Leenen, and H. S. Venter, "Social engineering attack framework," in *2014 Information Security for South Africa*. IEEE, 2014.

[11] A. Koyun and E. Al Janabi, "Social engineering attacks," *Journal of Multidisciplinary Engineering Science and Technology (JMEST)*, vol. 4, no. 6, pp. 7533–7538, 2017.

[12] K. Krombholz, H. Hobel, M. Huber, and E. Weippl, "Advanced social engineering attacks," *Journal of Information Security and applications*, vol. 22, 2015.

[13] P. Y. Leonov, A. V. Vorobyev, A. A. Ezhova, O. S. Kotelyanets, A. K. Zavalishina, and N. V. Morozov, "The main social engineering techniques aimed at hacking information systems," in *2021 Ural symposium on biomedical engineering, radioelectronics and information technology (USBEREIT)*. IEEE, 2021.

[14] R. Salama, F. Al-Turjman, S. Bhatla, and S. P. Yadav, "Social engineering attack types and prevention techniques-a survey," in *2023 International Conference on Computational Intelligence, Communication Technology and Networking (CICTN)*. IEEE, 2023.

[15] M. Volkamer, M. A. Sasse, and F. Boehm, "Analysing simulated phishing campaigns for staff," in *Computer Security: ESORICS 2020 International Workshops, DETIPS, DeSECSys, MPS, and SPOSE, Guildford, UK, September 17–18, 2020, Revised Selected Papers 25*. Springer, 2020.

[16] R. Marusenko, V. Sokolov, and P. Skladannyi, "Social engineering penetration testing in higher education institutions," in *International Conference on Computer Science, Engineering and Education Applications*. Springer, 2023.

[17] H. Aldawood and G. Skinner, "Reviewing cyber security social engineering training and awareness programs—pitfalls and ongoing issues," *Future internet*, vol. 11, no. 3, 2019.

[18] F. M. Teichmann and S. R. Boticiu, "An overview of the benefits, challenges, and legal aspects of penetration testing and red teaming," *International Cybersecurity Law Review*, vol. 4, no. 4, 2023.

[19] S. Labs, "Human interface device tutorial," 2011. [Online]. Available: https://www.silabs.com/documents/public/application-notes/AN249.pdf

[20] MIFARE, "Mifare official store." [Online]. Available: https://www.mifare.net/

Securing Authentication in Browser-Based Applications: Implementing an OAuth 2.0 Backend For Frontend compatible with Reverse Proxies

Alberto López-Trigo
Grupo de Ingeniería de Medios (GIM)
Cáceres, Spain
albertolt@unex.es

Fernando Calvino Balonero
Grupo de Ingeniería de Medios (GIM)
Cáceres, Spain
fcbalonero@unex.es

Agustín Javier Di Bartolo
Grupo de Ingeniería de Medios (GIM)
Cáceres, Spain
adibartolo@unex.es

Adrián Atienza Macías
Grupo de Ingeniería de Medios (GIM)
Cáceres, Spain
aatienza@unex.es

Mar Ávila Vegas
Grupo de Ingeniería de Medios (GIM)
Cáceres, Spain
mmavila@unex.es

***Abstract*—This article proposes an OAuth 2.0 Backend For Frontend (BFF) server component to secure authentication in browser-based applications. The architecture mitigates the security threats posed by token storage in a browser and token manipulation in a JavaScript environment by utilizing a server-side component to handle OAuth tokens. The implementation follows the Token Handler pattern and integrates with reverse proxies using the forward auth directive, ensuring secure token management outside the web browser. The OAuth BFF server component comprises an OAuth Agent, which interacts with the authorization server, and an OAuth Cookie Exchanger, which injects access tokens into requests. This solution, compliant with the "OAuth 2.0 for Browser-Based Applications" draft, offers scalability and seamless integration with existing web infrastructures.**

***Index Terms*—OAuth 2.0, Backend For Frontend, Web Security, OpenID Connect.**

I. Introduction

The use of the OAuth 2.0 protocol in browser-based applications can expose security threats due to the inherent risks of performing OAuth operations in a browser. One of the main problems is the lack of secure storage in a pure JavaScript environment. In a browser-based app, the only two alternatives to store access tokens are the `LocalStorage` API and the use of cookies. If the `LocalStorage` API is used as storage, tokens can be visible to third-party scripts (even from other domains) and vulnerable to known web attacks like Cross-Site Scripting (XSS) or Cross-Site Request Forgery (CSRF).One way to address this issue is to store the tokens in cookies and use the *HttpOnly* attribute, but between memory limitations, lack of control over when to send the token in an application, and loss of access to the token by the application, make this practice unfeasible and not recommended.

The Internet Engineering Task Force (IETF) has been working on different web application architectures that solve this problem, leading to the Active Internet-Draft called "OAuth 2.0 for Browser-Based Applications" [1]. This document proposes the *Backend For Frontend* (BFF) architecture, consisting of a server-side component responsible for interacting with the authorization server as a confidential OAuth client, managing and storing tokens in the context of a cookie-based session, and proxying all requests to the resource server, providing the correct access token.

The main contribution of this article is the implementation of a *Backend For Frontend* server component, as described in the Active Internet-Draft mentioned above. This implementation introduces three key innovations: (1) the design of a secure OAuth 2.0 token management mechanism that avoids storing tokens in the browser, (2) the integration of the *token handler* architecture proposed by Curity [2], allowing separation of concerns between the OAuth Agent and the Cookie Exchanger, and (3) seamless compatibility with widely used reverse proxies such as Traefik and NGINX through the *forward auth* directive. Additionally, the use of Redis for token storage allows scalability, load balancing, and deployment in distributed environments.

II. Proposed arquitecture

The OAuth BFF server component is divided into two separate functionalities: the OAuth Agent, responsible for interacting with the authorization server as a confidential OAuth client and managing the OAuth tokens, and the OAuth Cookie Exchanger, responsible for injecting the corresponding access token associated with the session cookie of the user.

A. OAuth Agent

The BFF OAuth Agent must obtain the OAuth tokens through the Authorization Code Flow as defined in the *RFC 6749* [3] using the Proof Key for Code Exchange (PKCE) extension proposed in *RFC 7636* [4]. Figure 1 represents

the OAuth 2.0 Authorization Code Grant using the PKCE extension carried out by the BFF OAuth agent.

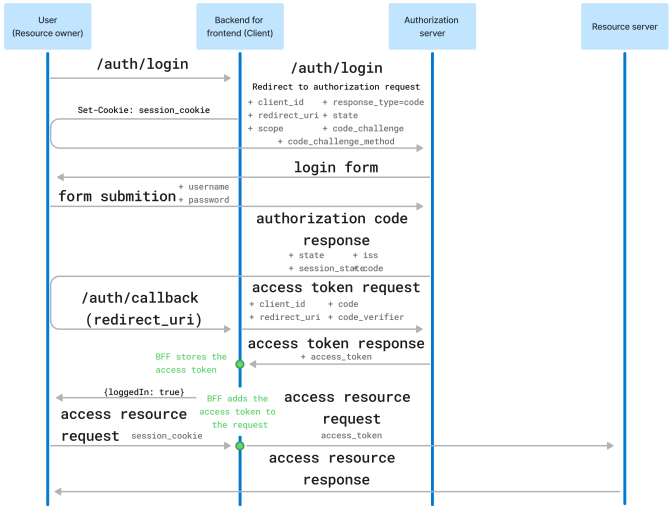


Fig. 1. OAuth 2.0 Authorization Code Flow with PKCE and BFF

The flow begins with an HTTP request to the BFF login endpoint. If the user has not logged in, the OAuth agent sets a session cookie to the user's browser and redirects the user to the Authorization server with the Authorization Code Grant and PKCE parameters. Once the OAuth agent exchanges the code for the access token (and more tokens in case of refresh tokens and OpenID Connect), the OAuth agent stores the token in a Redis database and notifies the user.

B. OAuth Cookie Exchanger

When the OAuth agent assigns a session cookie to the user and stored the access token in the Redis database, the OAuth Cookie Exchanger has the task of exchange the user's session cookie for the assigned session token and forwarding the HTTP request to the resource server. This is done via *forward auth* directives implemented on the reverse proxy. Figure 2 illustrates the process.

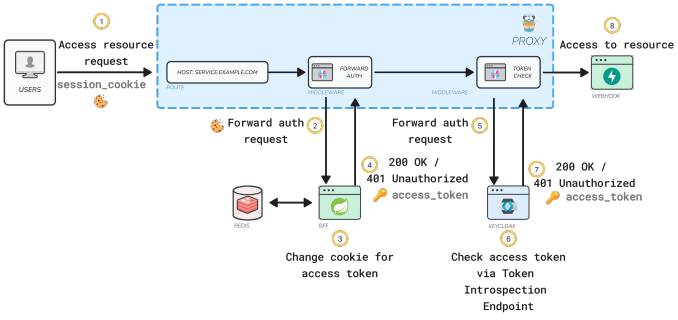


Fig. 2. OAuth Cookie Exchanger process

When the reverse proxy receives an HTTP request to a protected (1), it sends a copy of the request to the OAuth Cookie Exchanger using the *forward auth* directive and waits for a response (2). If the user has a cookie with an associated access token (3), the OAuth Cookie Exchanger then extracts the session cookie and queries the Redis database to retrieve the associated access token. If an access token is found, it injects the stored access token into the authorisation HTTP header and forwards the request with a `200 OK` response code (4). If the user does not have a token associated, the OAuth Cookie Exchanger responds with a `401 Unauthorized` response code and the proxy rejects the request.

If the proxy receives a `200 OK` response code, it sends the request with the access token to the service responsible for verifying the access token (5). This service can be the authorization server via the *OAuth 2.0 Token Introspection Endpoint* [5] or another service like *Open Policy Agen*t [6]. If the service verifies that the access token is valid (6), the HTTP request is approved (7), and the user receives the protected resource (8).

III. Conclusions

In summary, this article presents an OAuth *Backend For Frontend* server component that interfaces with an authorization server as a confidential OAuth client for web-based applications. The architecture mitigates risks of storing access tokens in the browser by delegating token management to a secure server-side component. Key contributions include: (1) a modular BFF design separating the OAuth Agent and Cookie Exchanger, (2) integration of the *Token Handler* pattern with Redis for scalable, stateless token storage, and (3) compatibility with reverse proxies like Traefik and NGINX via the *forward auth* directive.

This solution aligns with the IETF draft "OAuth 2.0 for Browser-Based Applications" and shows that secure token handling can be achieved without sacrificing scalability or integration.

Acknowledgment

This initiative is carried out within the framework of the funds of the Recovery, Transformation and Resilience Plan, financed by the European Union (Next Generation) - National Cybersecurity Institute (INCIBE) in the project C108/23 "Detection of Identity Document Forgery Using Computer Vision and Artificial Intelligence Techniques"

References

[1] A. Parecki, P. D. Ryck, and D. Waite, "OAuth 2.0 for Browser-Based Applications," Internet Engineering Task Force, Internet-Draft draft-ietf-oauth-browser-based-apps-24, Mar. 2025, work in Progress. [Online]. Available: https://datatracker.ietf.org/doc/draft-ietf-oauth-browser-based-apps/24/

[2] Curity, "Single Page Application Security," Curity, Whitepaper, 2024. [Online]. Available: https://curity.io/resources/documents/single-page-application-security-whitepaper/

[3] D. Hardt, "The OAuth 2.0 Authorization Framework," RFC 6749, Oct. 2012. [Online]. Available: https://www.rfc-editor.org/info/rfc6749

[4] N. Sakimura, J. Bradley, and N. Agarwal, "Proof Key for Code Exchange by OAuth Public Clients," RFC 7636, Sep. 2015. [Online]. Available: https://www.rfc-editor.org/info/rfc7636

[5] J. Richer, "OAuth 2.0 Token Introspection," RFC 7662, Oct. 2015. [Online]. Available: https://www.rfc-editor.org/info/rfc7662

[6] "Open policy agent — oauth2 and oidc samples." [Online]. Available: https://www.openpolicyagent.org/docs/latest/oauth-oidc/

Information Extraction and Homogeneity Validation of an Identity Document

Fernando Broncano Morgado
Grupo de Ingeniería de Medios
Universidad de Extremadura
fbroncano@unex.es

Marcos Jesús Sequera Fernández
Grupo de Ingeniería de Medios
Universidad de Extremadura
marcosjesus@unex.es

Sergio Guijarro Domínguez
Grupo de Ingeniería de Medios
Universidad de Extremadura
sergiosgd@unex.es

José Carlos Sancho Núñez
Grupo de Ingeniería de Medios
Universidad de Extremadura
jcsanchon@unex.es

***Abstract*—The digitisation of identity documents has evolved from manual processes to automated solutions, driven by advancements in computer vision. This study introduces a methodology for extracting and validating information from Spanish identity documents. The approach leverages YOLO for region detection, ORB for feature extraction, homography for image alignment with a template, and OCR for data interpretation. A validation system is also implemented to ensure consistency between the visual inspection zone –VIZ– and the machine readable zone –MRZ–, alongside check digits to confirm redundancies. Document homogeneity serves as a critical first step in verifying document authenticity.**

***Index Terms*—Identity document, digital image, image recognition, text recognition, authentication**

I. Introduction

The digitisation of physical documents into electronic information systems has always been a challenging task. Initially, document digitisation relied on manual processes involving the entry of information into the system. However, with advancements in computer vision, new mechanisms have emerged, enabling this manual task to transition into an automated process.

The development of automated information systems aims to facilitate the querying and storage of data. These systems are built on data models that enable structured information storage. Furthermore, they prevent issues such as duplication and verify data redundancy. Identity documents are a clear example of physical systems that require digitisation.

Current identity documents are still issued as physical documents, despite recent efforts to develop a digital version [1]. These documents contain relevant personal information that often needs to be processed mechanically. In response, the International Civil Aviation Organization proposed the creation of a standard document featuring a machine-readable zone [2]. Identity documents are structured into a visual inspection zone –VIZ–, which contains the document's information, and a machine readable zone –MRZ–, consisting of three lines of 30 OCR-B characters designed to be easily interpreted by machines.

Various organizations require systems for the mechanization of identity documents. In this regard, several studies [3], [4] have been proposed in the literature that address automatic reading and information recognition in identity documents to meet this need.

Tools for the mechanization of physical information must extract and process data from digital images. Traditionally, computer vision systems have relied on algorithms such as SURF [5], SIFT [6], and ORB [7] to study key points in images, or optical character recognition –OCR– platforms for character extraction. Additionally, with the rise of artificial intelligence and the introduction of more advanced models, solutions like *You Only Look Once* –YOLO– [8] have been proposed, enabling the detection of objects within an image.

This work proposes the development of a methodology for extracting information from a Spanish identity document. Additionally, it introduces a process for verifying the homogeneity of a document by comparing the machine readable zone –MRZ– with the visual inspection zone –VIZ–.

II. Data Extraction from an Identity Document

Traditionally, reading a document has involved identifying the region of interest within an image and applying a character recognition algorithm to interpret its meaning. Based on this approach, a methodology for extracting information from an identity document is proposed, leveraging these characteristics.

The proposed methodology for extracting data from an identity document consists of the following steps: (1) identifying the region of interest containing the identity document within an image; (2) extracting features using ORB from that region of interest; (3) applying a homography process to map key points between the region of interest and the base template; (4) cropping the regions of interest according to a table of positions and sizes; (5) applying thresholding and optical character recognition to extract the data associated with that

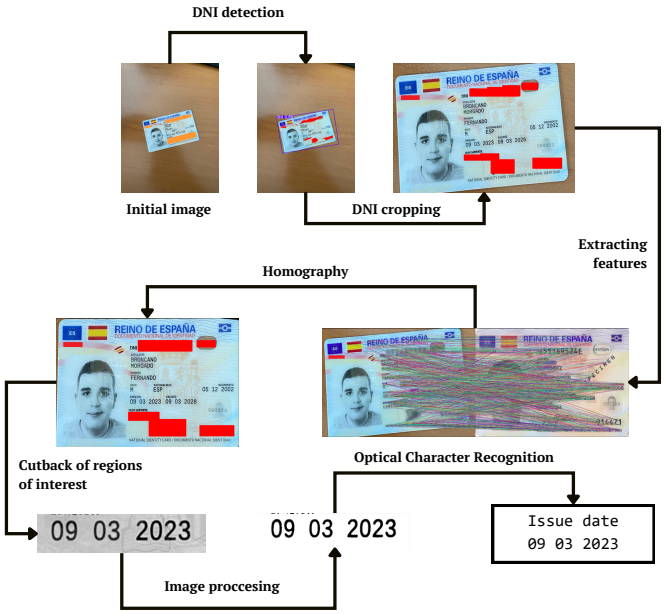


Figure 1. Information extraction using the proposed methodology

region of interest. This methodology is illustrated graphically in Figure 1.

Prior to applying this methodology, a template of an identity document must be selected. A table of regions of interest should then be created for this template and linked to the corresponding data. This table is used to perform cropping operations on the provided image. Additionally, a *You Only Look Once* –YOLO– model with oriented bounding boxes capabilities has been trained to identify the region containing the document with maximum precision.

III. Homogeneity of Information in an Identity Document

Once the information from an identity document has been extracted, the existing redundancy can be verified using regular expressions and information redundancy checks. The primary redundancy feature in an identity document is the cross-checking of information between the visual inspection zone –VIZ– and the machine readable zone –MRZ–. An example of a machine-readable zone can be seen in Figure 2.

The verification of this information is performed once the data has been extracted and structured according to its corresponding feature. The information from both zones must match. Additionally, the Spanish National Identity Document –DNI– includes a redundancy mechanism through a security letter in the DNI number, as well as a consistency check between the issuance date, date of birth, and validity period.

Within the machine readable zone –MRZ–, several check

Figure 2. Example of a machine readable zone

digits –CD– are present. These check digits are calculated and compared with those present in the MRZ. In this way, redundancy within the machine-readable zone is verified.

IV. Conclusions

This work proposes a methodology for extracting information from an identity document using YOLO, ORB, and OCR. Additionally, the validity of the information is verified by cross-checking redundancy and ensuring homogeneity between zones. As future work, advancements should focus on expanding the dataset used for region-of-interest recognition, as well as improving base templates and interest zone tables.

Acknowledgements

This initiative is carried out within the framework of the funds of the Recovery, Transformation and Resilience Plan, financed by the European Union –Next Generation– and National Cybersecurity Institute –INCIBE– in the Project C108/23 “Detección de Falsificación de Documentos de Identidad mediante Técnicas de Visión por Computador e Inteligencia Artificial”.

References

[1] *Real Decreto 255/2025, de 1 de abril, por el que se regula el Documento Nacional de Identidad.*, Boletín Oficial del Estado, 2025. [Online]. Available: https://www.boe.es/eli/es/rd/2025/04/01/255/

[2] *Doc 9303 Documentos de viaje de lectura mecánica*, Organización de la Aviación Civil Internacional, 2021. [Online]. Available: https://www.icao.int/publications/pages/publication.aspx?docnum=9303

[3] S. Carta, A. Giuliani, L. Piano, and S. G. Tiddia, “An end-to-end ocr-free solution for identity document information extraction,” in *Procedia Computer Science*, vol. 246, 2024, p. 453 – 462.

[4] M. K. Gupta, R. Shah, J. Rathod, and A. Kumar, “Smartidocr: Automatic detection and recognition of identity card number using deep networks,” in *Proceedings of the IEEE International Conference Image Information Processing*, vol. 2021-November, 2021, p. 267 – 272.

[5] H. Bay, T. Tuytelaars, and L. Van Gool, “Surf: Speeded up robust features,” in *Computer Vision – ECCV 2006*, 2006, pp. 404–417.

[6] D. Lowe, “Object recognition from local scale-invariant features,” in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, 1999, pp. 1150–1157 vol.2.

[7] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, “Orb: An efficient alternative to sift or surf,” in *2011 International Conference on Computer Vision*, 2011, pp. 2564–2571.

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

Zero-Query Black-box Adversarial Attack

Joana Cabral Costa
sins-lab, Instituto de Telecomunicações
Universidade da Beira Interior
Covilhã, Portugal
joana.cabral.costa@ubi.pt

Tiago Roxo
sins-lab, Instituto de Telecomunicações
Universidade da Beira Interior
Covilhã, Portugal
tiago.roxo@ubi.pt

Hugo Proença
sins-lab, Instituto de Telecomunicações
Universidade da Beira Interior
Covilhã, Portugal
hugomcp@ubi.pt

Pedro R. M. Inácio
sins-lab, Instituto de Telecomunicações
Universidade da Beira Interior
Covilhã, Portugal
prmi@ubi.pt

***Abstract*—The existence of adversarial samples is a threat to the application of Deep Learning models in critical areas. To generate these, there are white-box attacks that directly access model gradients and loss functions and black-box attacks that query the target model a significant number of times to create adequate adversarial perturbations. However, the possibility of performing an attack without querying the model is still little explored. Therefore, we propose a Zero-Query Black-Box Adversarial (ZQBA) attack that uses residual features of images to create adversarial samples, significantly reducing model performance without querying the target model or accessing model gradients.**

***Index Terms*—adversarial attack, gray-box, residual image, zero-query**

I. Introduction

Deep Learning models are susceptible to imperceptible image perturbations that compromise their performance, typically denominated by adversarial samples, which is critical in security-related areas. These samples can derive from spontaneous perturbations (*e.g.*, noise), but can also be created maliciously via adversarial attacks, commonly grouped into white-box and black-box attacks. Contrary to existing approaches that require access or querying the attacked model, we propose a Zero-Query Black-box Adversarial (ZQBA) attack that creates adversarial perturbations by combining residual features of images from different classes, significantly reducing the attacked model performance.

II. Related Work

White-box Attacks. White-box methods typically rely on assessing model gradients to create adversarial samples specifically suited to attack the model. Fast Gradient Sign Method (FGSM) [1] is a one-step method that finds adversarial samples using the model loss function. Projected Gradient Descent (PGD) [2] uses saddle point formulation to find a strong perturbation through multiple iterations. Auto-Attack [3] combines four attacks (the majority are white-box) to duly evaluate the robustness of a defense.

Black-box Attacks. Black-box attacks usually require multiple queries from the target models to create a suitable adversarial sample. Square Attack [4] uses a randomized search scheme that perturbs the images in localized square-shaped in random positions. Disentangled Feature Space (DifAttack) [5] trains an autoencoder on clean and adversarial examples to generate strong perturbations against the victim model. Park *et al.* [6] propose a practical way of using hard-label-based attacks, using a surrogate model, achieving higher query efficiency.

Our Approach. Relative to existing methods, our approach can create adversarial samples to disrupt model performance without any queries (an improvement over black-box settings), and it is only based on the knowledge of the architecture and training data of the model to attack, *i.e.*, without requiring access to gradients of the attacked model (an improvement over white-box settings).

III. Methodology and Preliminary Results

Models, Dataset and Evaluation Metrics. For models, we use ResNet-18, ResNet-50, and MobileNet-v2 from the publicly available PyTorch [7] library, while for performance evaluation we use accuracy on natural and adversarial test samples, using CIFAR-10 data [8]. To evaluate the quality of the generated images, we use the Peak Signal-to-noise Ratio (PSNR), as previously done in [9].

Obtaining Residual Feature. To obtain residual features, we use Guided BackPropagation [10], which is a gradient-based approach that retrieves the gradient of images when backpropagating through the Rectified Linear Unit (ReLU) activation function, where only the flow of positive gradients is allowed by changing the negative gradient values to zero. This is typically used to display the features of the input image that maximized the activation of the feature maps, meaning that it more closely influenced the model prediction. For ResNet18 and ResNet50, this information was extracted from the 4th

This work was supported in part by the Portuguese Fundação para a Ciência e Tecnologia (FCT)/Ministério da Ciência, Tecnologia e Ensino Superior (MCTES) through National Funds and co-funded by EU funds under Project UIDB/50008/2020; in part by the FCT Doctoral Grant 2021.04905.BD.

TABLE I
PERFORMANCE OF MULTIPLE ARCHITECTURES ON THE ZQBA ATTACK USING DIFFERENT BASELINES TO CREATE RESIDUAL FEATURE, ON CIFAR-10. PAR(M) REFERS TO THE NUMBER OF PARAMETERS IN MILLIONS.

Target Model	Par(M)	Clean	ZQBA		
			R50	R18	MN2
ResNet50	22.4	96.65	62.66	60.11	53.11
ResNet18	10.7	94.43	63.55	59.32	48.86
MobileNetv2	2.1	85.08	57.79	54.24	35.57

layer, and for MobileNetv2, it was extracted from the 18th feature layer.

Residual Feature Performance Effect. We start by training different models in CIFAR-10 to retrieve residual features of each dataset image via Guided BackPropagation. Then, we assess the effect of said residual features on model performance when applying them in different CIFAR-10 images in Table I. Note that the testing models are not the same as the ones used to retrieve residual features, although they share the same architecture and training data. At this stage, our approach to disrupt model predictions was to include random residual features, resulting in a significant decrease in model performance relative to clean data, with the most considerable decrease when applying residual features from architectures with fewer parameters. In smaller models, the residual feature may be more representative, contributing to more disruptive performances when applied to images, regardless of the model attacked.

Residual Feature Perturbation. The inclusion of residual features to CIFAR-10 images could potentially result in significant visual disruption of the image, which is not typically expected from adversarial attacks [11]. As such, we assess the visual effect of these inclusions by comparing the original image (and its respective residual feature) with the target image and the resulting combination in Figure 1. The results show that the Attacked Image does not significantly change relative to its original state, as visually shown, and given their PNSR values, highlighting the relevance of our approach for zero-query adversarial attacks.

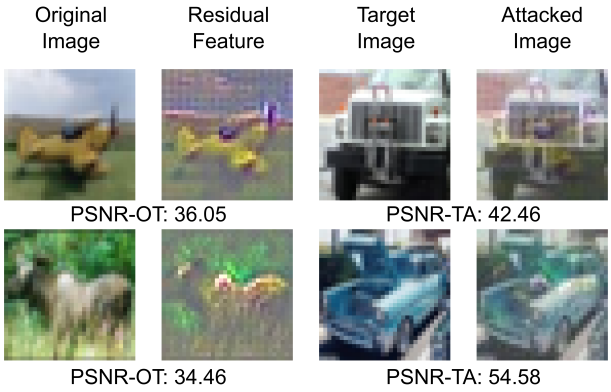


Fig. 1. Original Image, Residual Feature obtained from the original image, Target Image, and Attacked Image obtained by combining the Residual Feature with Target Image for CIFAR-10 dataset. PSNR-OT means PSNR between Original and Target images, and PSNR-TA means PSNR between Target and Attacked images. Higer PNSR relates to increased image similarity.

IV. CONCLUSION AND FUTURE DIRECTIONS

We propose a Zero-Query Black-box Adversarial attack that, without querying the target model, significantly reduces the model performance. The preliminary results highlight the applicability of residual features for zero-query attacks, but only the setup of random residual feature inclusion on CIFAR-10 images was evaluated to disrupt models performance. As such, we have some directions to explore for our future work:

- Include residual features from similar images (but different classes) onto the target image to assess the effect of the attack using our approach. The similarity of images could be retrieved using the Structural Similarity Index, the Information theoretic-based Statistic Similarity Measure, or transformer-based models [12];
- Complement the experiments with defense approaches to assess the resilience of our zero-query approach relative to state-of-the-art defenses, namely, Adversarial Training, Diffuse Purification, or Adversarial Distillation;
- Extend the results of Table I to other architectures and datasets, namely CIFAR-100 and Tiny ImageNet, to assess the robustness of our approach to various image conditions, and extend the results by comparing with black-box attacks using no queries, to more accurately assess the strength of our attack in a zero-query setup.

REFERENCES

[1] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[2] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *ICLR*, 2018.

[3] F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in *ICML*, pp. 2206–2216, PMLR, 2020.

[4] M. Andriushchenko, F. Croce, N. Flammarion, and M. Hein, "Square attack: a query-efficient black-box adversarial attack via random search," in *ECCV*, pp. 484–501, Springer, 2020.

[5] J. Liu, J. Zhou, J. Zeng, and J. Tian, "Difattack: Query-efficient black-box adversarial attack via disentangled feature space," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 3666–3674, 2024.

[6] J. Park, P. Miller, and N. McLaughlin, "Hard-label based small query black-box adversarial attack," in *Proceedings of the IEEE/CVF WACV*, pp. 3986–3995, 2024.

[7] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," *Open Review*, 2017.

[8] A. Krizhevsky, G. Hinton, *et al.*, "Learning multiple layers of features from tiny images," *Master's thesis, University of Tront*, 2009.

[9] T. Bui, S. Agarwal, N. Yu, and J. Collomosse, "Rosteals: Robust steganography using autoencoder latent space," in *Proceedings of the IEEE/CVF conference on CVPR*, pp. 933–942, 2023.

[10] S. Mostafa, D. Mondal, M. A. Beck, C. P. Bidinosti, C. J. Henry, and I. Stavness, "Leveraging guided backpropagation to select convolutional neural networks for plant classification," *Frontiers in Artificial Intelligence*, vol. 5, p. 871162, 2022.

[11] F. Croce, M. Andriushchenko, V. Sehwag, E. Debenedetti, N. Flammarion, M. Chiang, P. Mittal, and M. Hein, "Robustbench: a standardized adversarial robustness benchmark," *arXiv preprint arXiv:2010.09670*, 2020.

[12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, pp. 8748–8763, PmLR, 2021.

Intelligent and Cybersecure Management of Construction and Demolition Waste by Digital Images

Ruth Torres Gallego
Grupo de Ingeniería de Medios (GIM)
University of Extremadura
Cáceres, Spain
rutorresg@alumnos.unex.es

Pablo García Rodríguez
Grupo de Ingeniería de Medios (GIM)
University of Extremadura
Cáceres, Spain
pablogr@unex.es

Jesús Torrecilla Pinero
Grupo Benito Mahedero
University of Extremadura
Cáceres, Spain
jtorreci@unex.es

Aurora Cuartero Sáez
Grupo de investigación Kraken
University of Extremadura
Cáceres, Spain
acuartero@unex.es

Pablo Plaza Caballero
Grupo Benito Mahedero
University of Extremadura
Cáceres, Spain
pablopc@unex.es

***Abstract*—This paper presents an innovative system for the automatic identification of trucks, qualitative characterization, and quantification of construction and demolition waste (CDW) at treatment plants using digital images with artificial vision and machine learning techniques. Additionally, it addresses the relevance of cybersecurity in protecting generated data, ensuring integrity and confidentiality in a critical industrial environment.**

***Keywords*—Construction and demolition waste (CDW), digital image, pattern recognition and image analysis, artificial vision, machine learning, cybersecurity**

I. Introduction

Efficient Construction and Demolition Waste (CDW) management is a global challenge due to its volume and heterogeneity. Traditional methods, based on manual inspections and physical weighing, are slow and error-prone. This project proposes an automated system using surveillance cameras to perform several tasks described in the following paragraphs [1].

Firstly, identify trucks using artificial vision techniques through their license plate (Fig. 1), and in this way open a file in the database with that truck and other data of interest that in many cases may be previously covered to give access [2].

Secondly, classify waste materials (e.g., wood, concrete, plastic) using machine learning and/or deep learning algorithms (CNNs) applied to video streams or still images, enabling accurate vehicle detection, classification, and tracking within construction and demolition waste management environments [3].

Finally, the payload volume needs to be calculated using 3D vision algorithms and correlated with densities derived from historical weighing data [4].

In parallel, cybersecurity protocols are integrated to protect infrastructure against cyberattacks that could disrupt operations or manipulate critical data [5].

II. Methodology

The proposed system integrates sequential workflows for waste management and cybersecurity. The first three steps focus on data acquisition and analysis and the forth step is in relation with cybersecurity.

Fig. 1. A truck with CDW

A. Truck Identification

Surveillance cameras capture real-time footage of incoming trucks. YOLOv5, a state-of-the-art object detection model, identifies trucks by their shape and license plates, enabling automated entry logging. This step ensures traceability and links waste loads to specific vehicles (Fig. 2).

B. Waste Characterization

Once a truck is detected, a CNN model analyzes high-resolution images of the waste payload. The model, trained on a proprietary dataset of labeled CDW images (e.g., concrete, wood, metal), classifies materials using texture, color, and spatial features (Fig. 3). Physical samples validate the model's accuracy, addressing variability in waste composition.

C. Quantification

Stereoscopic cameras generate 3D point clouds of the truck's payload to estimate volume. Historical density data (e.g., concrete = 2.4 t/m³, wood = 0.8 t/m³) is then correlated with volumetric measurements via linear regression, enabling weight estimation without physical scales (Fig. 4).

D. Cibersecurity

The fourth step, cybersecurity, is critical to safeguarding the integrity and reliability of the entire system. As the initial stages generate sensitive operational data—including truck identifiers, waste volume measurements, and material classifications—the system becomes a potential target for cyberattacks. Without proper protection, malicious actors could tamper with identification mechanisms (e.g., spoof license plates to circumvent access controls), manipulate classification outputs (e.g., mislabel hazardous materials as inert), or falsify weight and volume estimations (e.g., inflate

payloads to commit billing fraud), ultimately undermining both operational and regulatory objectives [6].

To mitigate these threats, a comprehensive cybersecurity strategy is implemented. Data transmissions are protected through robust encryption protocols, while multi-factor authentication restricts unauthorized access to the system. Additionally, an intrusion detection system based on Snort continuously monitors network traffic for anomalies and known attack patterns. Together, these mechanisms ensure the confidentiality, integrity, and availability of the system, supporting compliance with environmental regulations and minimizing the risk of service disruptions caused by cyber intrusions.

Fig. 2. A truck unloading material (CDW)

III. Expected Results

The proposed system is expected to deliver high-performance results across multiple operational dimensions. Preliminary evaluations indicate a truck identification accuracy of approximately 98%, while material classification achieves an accuracy rate of 92% across various construction and demolition waste types. Volume estimation exhibits a deviation of less than 5% when compared to manual measurement benchmarks, demonstrating strong reliability in automated assessments. Additionally, the integration of intelligent processing pipelines contributes to a 30% reduction in overall processing time. From a cybersecurity perspective, the system has successfully withstood simulated attack scenarios—including false data injection—thereby validating the effectiveness of the implemented protection mechanisms.

Fig. 3. Different CDW to classify

IV. Discussion

The system optimizes CDW logistics, promoting a circular economy. Cybersecurity emerges as a critical pillar: attacks could falsify weighing data (impacting billing) or paralyze operations by contaminating waste stockpiles. Combining computer vision with robust security measures ensures reliability and scalability.

V. Conclusions

This work demonstrates the feasibility of automating CDW management through AI techniques, highlighting the need to integrate cybersecurity from the design phase. Future research will expand to other waste types and explore blockchain for data auditing.

Fig. 4. Processed CDW in the lab (GARNOCEX project) [7]

Acknowledgments

This initiative is carried out within the framework of the funds of the Recovery, Transformation and Resilience Plan, financed by the European Union (Next Generation) - National Cybersecurity Institute (INCIBE) in the project C107/23 “Artificial Intelligence Applied to Cybersecurity in Critical Water and Sanitation Infrastructures”.

References

[1] Mhlanga, D., & Shao, D. (2025). AI-optimized urban resource management for sustainable smart cities. In Financial Inclusion and Sustainable Development in Sub-Saharan Africa (pp. 96–116). Routledge.

[2] Lu, W., & Chen, J. (2022). Computer vision for solid waste sorting: A critical review of academic research. Waste Management, 142 , 29–43.

[3] Lin, K., Zhou, T., Gao, X., Li, Z., Duan, H., Wu, H., ... & Zhao, Y. (2022). Deep convolutional neural networks for construction and demolition waste classification: VGGNet structures, cyclical learning rate, and knowledge transfer. Journal of Environmental Management, 318 , 115501.

[4] Chen, J., Lu, W., Yuan, L., Wu, Y., & Xue, F. (2022). Estimating construction waste truck payload volume using monocular vision. Resources, Conservation and Recycling, 177 , 106013.

[5] Humayed, A., Lin, J., Li, F., & Luo, B. (2017). Cyber-physical systems security—A survey. IEEE Internet of Things Journal, 4 (6), 1802–1831.

[6] Brighente, A., Conti, M., Di Renzone, G., Peruzzi, G., & Pozzebon, A. (2023). Security and privacy of smart waste management systems: A cyber–physical system perspective. IEEE Internet of Things Journal, 11 (5), 7309–7324.

[7] GARNOCEX, Research project for the use of recycled aggregates in road infrastructure in Extremadura: https://www.garnocex.es/

Satellite Image-Based Water Quality Index Maps. Data Cybersecurity

V. Amores-Chaparro
Universidad de Extremadura
vicamoresc@unex.es

A. Cuartero
Universidad de Extremadura
acuartero@unex.es

J. Torrecilla
Universidad de Extremadura
jtorreci@unex.es

***Abstract*—This work presents a satellite-based approach for monitoring water quality parameters—chlorophyll, turbidity, and suspended solids—using spectral indices from multispectral imagery. The method supports large-scale, cost-effective analysis. However, integrating satellite data processing with cloud platforms introduces cybersecurity risks. A verification method is proposed to improve data integrity.**

***Index Terms*—Cloud Computing, Apache Spark, Remote Sensing, Sentinel-2, Copernicus.**

I. Introduction

MONITORING water quality at scale is vital for ecosystem health and safe drinking water. Traditional in-situ methods are limited in coverage and frequency, while remote sensing enables scalable analysis using satellite-derived spectral indices. Key examples include NDCI for chlorophyll-a, red/green reflectance for turbidity, and SWIR bands for suspended solids (TSM). However, these systems face cybersecurity risks such as data traceability issues and processing vulnerabilities. This study proposes a secure satellite-based framework combining spectral analysis with cybersecurity measures to ensure data integrity [1].

II. Methodology

This study relies on the acquisition and preprocessing of high-quality satellite imagery, primarily obtained from the Sentinel-2 satellite within the Copernicus program of the European Union. Additional data from platforms such as Landsat-8 and MODIS can be incorporated to enhance spatial and temporal coverage.

A. Satellite Image Acquisition and Preprocessing

1) Satellite Image Acquisition: For the experiments described in this study, the SentinelHub platform API [1] was utilized. In particular, the images used in this study correspond to the area of The Molinos reservoir, located in the province of Badajoz. These images, obtained via the Sentinel-2 satellite, provide a detailed view of the region by leveraging its 13 different spectral bands. The focus on The Molinos reservoir allows for the validation of remote sensing index results with situ experimental measurements (see Fig. 1).

Sentinel-2's 13 spectral bands—from visible to shortwave infrared—enable the calculation of indices for water quality analysis. The visible bands (B2-B4) estimate turbidity and solids; the red band (B5-B7) supports chlorophyll a estimation (e.g. NDCI); and B8, B11, B12 are used for TSM and detection of algae bloom.

[1]https://www.sentinel-hub.com/

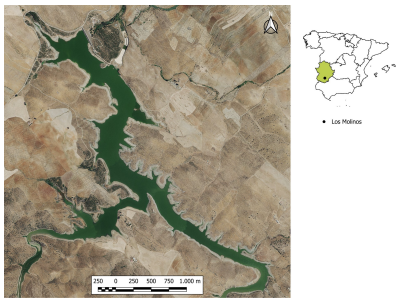


Fig. 1. The study area: The Molinos reservoir, in the Southwest of Spain.

2) Preprocessing Steps: To ensure reliable water quality indices, the following preprocessing steps are applied:

- Image Acquisition: Sentinel-2 imagery is obtained via platforms like Copernicus Open Access Hub.
- Atmospheric Correction: Tools such as Sen2Cor or ACOLITE remove atmospheric effects (e.g., aerosols, water vapor), ensuring accurate surface reflectance [2], [3].
- Cropping and Reprojection: Images are clipped to the study area and reprojected to a common spatial reference.
- Cloud Masking: Cloudy pixels are excluded using automated detection (e.g., Sen2Cor).
- Radiometric calibration: Digital numbers are converted to TOA or surface reflectance for consistent temporal analysis.

B. Spectral Index Calculation

Spectral indices derived from satellite images employ reflectance values in different bands. These indices serve as indirect indicators of water quality parameters such as chlorophyll-a concentration, turbidity, and total suspended solids (TSM). Calculations use the Apache Spark library, ensuring computational efficiency [4]. All these index maps were processed over two consecutive months of the hottest season (June and July) and from two different years (2023 and 2024).

1) Chlorophyll-a Estimation: Chlorophyll-a concentration is estimated using the Normalized Difference Chlorophyll Index (NDCI) [5], which exploits the spectral signature of chlorophyll in the red-edge and near-infrared regions. NDCI (Chlorophyll-a): $(R_{783}-R_{705})/(R_{783}+R_{705})$, Figure 2 shows the Normalized Difference Water Index Map.

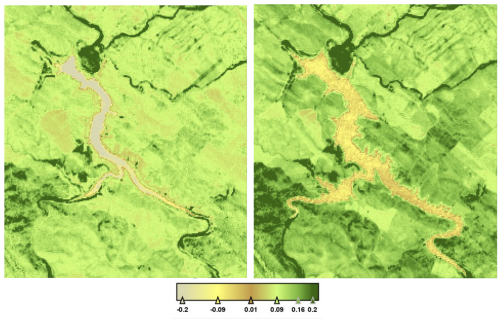


Fig. 2. Normalized Difference Chlorophyll Index Maps (NDCI) Maps in Los Molinos Reservoir: July 2023 (le ft) and June 2024 (right)

2) Turbidity Estimation: Turbidity is estimated using two alternative methods, either by using reflectance in the red band (R_{665}), which correlates strongly with light scattering caused by suspended particles or by using normalized index NDTI. NDTI (Turbidity): $(R_{665}-R_{560})/(R_{665}+R_{560})$ Figure 3 show the normalized turbidity index map.

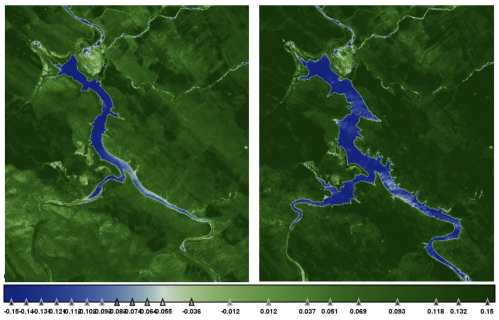


Fig. 3. Normaliced Turbity index Maps in Los Molinos Reservoir: July 2023 (left) and June 2024 (right)

3) Suspended Solids Quantification: Total Suspended Solids (TSM) are estimated using SWIR-based algorithms, including custom and standard indices like the Normalized Difference Water Index (NDWI) [6]. Figure 4 shows maps of the suspended solids index.

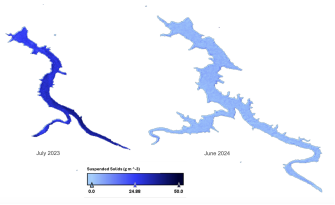


Fig. 4. Suspended Solid index Maps in The Molinos reservoir: July 2023 (left) and June 2024 (right)

4) Integration with Spark Library Algorithms: Apache Spark enables scalable processing of large satellite datasets through:

- Distributed Computation: RDDs allow parallel spectral index calculation across millions of pixels, significantly reducing runtime.
- ML Integration: Spark MLlib supports training models (e.g., Random Forest) to enhance index–parameter correlations,and enhance regions of interest [7], [8].
- Fault Tolerance: Spark's architecture ensures reliable handling of large-scale data.

This integration supports fast, accurate water quality estimation, facilitating real-time environmental monitoring.

By combining robust spectral index calculations with scalable Spark-based processing, this study achieves accurate and efficient e stimation o f w ater q uality p arameters, p aving the way for real-time environmental monitoring.

III. Conclusion and Future Work

This study proposes the effectiveness of combining multispectral imaging, Apache Spark, and cybersecurity validation for water quality monitoring. Future work will focus on real-time index validation and threat detection automation.

Acknowledgment

This initiative is carried out within the framework of the funds of the Recovery, Transformation and Resilience Plan, financed by the European Union (Next Generation) - National Cybersecurity Institute (INCIBE) in the project C107/23 "Artificial Intelligence Applied to Cybersecurity in Critical Water and Sanitation Infrastructures".

References

[1] M. Homaei, A. Caro Lindo, O. Mogollon-Gutierrez, and J. Diaz Alonso, "The role of artificial intelligence in digital twin's cybersecurity," in *XVII Reunión Española sobre Criptología y Seguridad de la Información (RECSI)*, vol. 265, 2022, p. 133. [Online]. Available: https://doi.org/10.22429/Euc2022.028

[2] J. Cáceres Merino, A. Cuartero Sáez, and J. Á. Torrecilla Pinero, "Finding optimal spatial window: the influence of size on remote-sensing-based chl-a prediction in small reservoirs," *IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING*, vol. 17, p. 18769, 2024.

[3] A. Cuartero, J. Cáceres-Merino, and J. A. Torrecilla-Pinero, "An application of c2-net atmospheric corrections for chlorophyll-a estimation in small reservoirs," *Remote Sensing Applications: Society and Environment*, vol. 32, p. 101021, 2023.

[4] M. Saberioon, J. Brom, V. Nedbal, P. Soucek, and P. Císar, "Chlorophyll-a and total suspended solids retrieval and mapping using sentinel-2a and machine learning for inland waters," *Ecological Indicators*, vol. 113, p. 106236, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1470160X20301734

[5] S. Mishra and D. R. Mishra, "Normalized difference chlorophyll index: A novel model for remote estimation of chlorophyll-a concentration in turbid productive waters," *Remote Sensing of Environment*, vol. 117, pp. 394–406, 2012, remote Sensing of Urban Environments. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0034425711003737

[6] Y. Du, Y. Zhang, F. Ling, Q. Wang, W. Li, and X. Li, "Water bodies' mapping from sentinel-2 imagery with modified normalized difference water index at 10-m spatial resolution produced by sharpening the swir band," *Remote Sensing*, vol. 8, no. 4, 2016. [Online]. Available: https://www.mdpi.com/2072-4292/8/4/354

[7] R. Molano, D. Caballero, P. G. Rodríguez, M. D. M. Ávila, J. P. Torres, M. L. Durán, J. C. Sancho, and A. Caro, "Finding the largest volume parallelepipedon of arbitrary orientation in a solid," *IEEE Access*, vol. 9, pp. 103 600–103 609, 2021.

[8] R. Molano, M. Ávila, J. C. Sancho, P. G. Rodríguez, and A. Caro, "An algorithm to compute any simple k-gon of a maximum area or perimeter inscribed in a region of interest," *SIAM Journal on Imaging Sciences*, vol. 15, no. 4, pp. 1808–1832, 2022.

Classification and analysis of LiDAR data

Pablo Fernández González
Grupo de Ingeniería de Medios (GIM)
University of Extremadura
Cáceres, Spain
pfernandzoq@alumnos.unex.es

Pablo García Rodríguez
Grupo de Ingeniería de Medios (GIM)
University of Extremadura
Cáceres, Spain
pablogr@unex.es

Aurora Cuartero Sáez
Grupo de investigación Kraken
University of Extremadura
Cáceres, Spain
acuartero@unex.es

Victoria Amores Chaparro
Grupo de Ingeniería de Medios (GIM)
University of Extremadura
Cáceres, Spain
vicamoresc@unex.es

Fernando Broncano Morgado
Grupo de Ingeniería de Medios (GIM)
University of Extremadura
Cáceres, Spain
fbroncano@unex.es

***Abstract*—In recent times, LiDAR technology has seen a steady rise in popularity, mostly due to its broad use in autonomous navigation, environmental mapping, and urban planning applications. Despite its growing adoption, it still lacks a standardized methodological framework that facilitates its deployment with these types of data, in order to improve precision when classifying clusters of points and comparing z-axis values. Therefore, a study will be conducted comparing different applications aimed at controlling accuracy levels to better support this data engineering process. Once the data is properly structured, its metadata, along with classification algorithms, will be key in optimizing workflows. Ultimately, deployments will be designed and assessed through real-life examples, establishing objective methodological steps and identifying future real-world applications in the process joining Artificial Intelligence techniques with cybersecurity-enhanced algorithms.**

Keywords—LiDAR, data engineering, classification algorithms, cibersecurity

I. Introduction

Light Detection and Ranging (better known as *LiDAR*) is a remote sensing technology, based on the use of laser light to measure distances between a sensor and objects in its environment.

By emitting laser pulses and analyzing the time it takes for them to return after reflecting off surfaces, LiDAR systems are able to precisely generate three-dimensional representations of the scanned area, also known as point clouds. These consist of numerous data points with x, y, and z coordinates, capturing the spatial structure of the environment [1].

LiDAR data processing usually involves several key steps to transform raw point cloud data into usable information: Initially, the data undergoes preprocessing to correct any errors and align it with geographic coordinate systems. Subsequent steps include filtering to remove noise, classification to differentiate between ground and non-ground points (or other possible values, such as vegetation or bodies of water), and the generation of digital models such as Digital Terrain Models (*DTMs*) and Digital Surface Models (*DSMs*) [2].

Its role in cybersecurity, however, should not be undermined, as LiDAR technology is increasingly integral to securing cyber-physical systems by providing precise, real-time 3D spatial data. LiDAR enhances perimeter surveillance and intrusion detection across various sectors, including critical infrastructure, data centers, and automated industrial facilities. Furthermore, its ability to properly operate under diverse weather and lighting conditions ensures continuous monitoring and reduces false alarms, strengthening physical security measures as a result [3, 4].

However, despite its widespread use in fields such as autonomous navigation, or environmental mapping, there remains a lack of standardized methodologies for processing and analyzing LiDAR data; a gap that can lead to inconsistencies in data quality and challenges in integrating LiDAR-derived information across different applications. Thus, developing standardized frameworks and best practices for LiDAR data processing is crucial to enhance the accuracy and reliability of analyses, particularly when classifying point clusters and comparing elevation values along the z-axis.

Thus, the main goal of this project is to compare the classification of LiDAR data using various real datasets or alternative technologies.

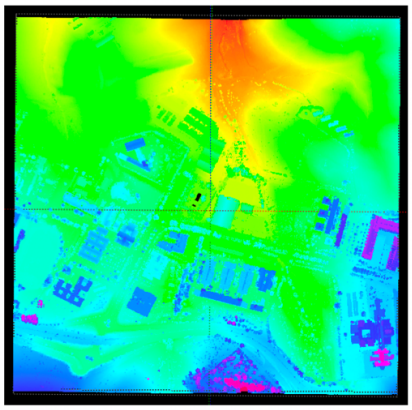


Fig. 1. Digital Elevation Model (DEM) Visualization with Z-Axis Elevation Data – Color-Encoded Topography of an Urban and Terrain Area

II. Previous research

Early investigations into LiDAR technology laid the groundwork for its application across various disciplines. While initial studies focused on developing geometric and statistical approaches for processing point cloud data -such as methods based on Triangular Irregular Networks (TIN) to reconstruct digital terrain models (DTMs) and separate ground points from non-ground features [5] [1], as technology

[1] This period also saw the establishment of standardized file formats and classification schemes (e.g., the **LAS** format [6], as in *LASer*) that provided a common basis for data exchange and subsequent processing workflows.

matured, researchers began addressing data noisiness and the inherent variability in spatial sampling.

Early classification approaches were primarily rule-based, relying on user-defined thresholds and geometric properties (such as z-axis elevation, intensity, and return numbers) to classify point clouds. Software tools like TerraScan [7] became critical by automating the classification process through a combination of macro-driven algorithms and manual quality control, thereby providing practical solutions to filter ground points from vegetation and urban structures.

Machine learning, later on, would further transform LiDAR data processing. Seminal studies applied both supervised and unsupervised learning techniques to improve classification accuracy and robustness [8]. Furthermore, deep learning methodologies are currently at the forefront of LiDAR research, as pioneering neural network architectures enabled the direct ingestion of raw point cloud data, facilitating simultaneous extraction of both local and global features. Convolutional Neural Networks (*CNNs*) and transformer-based models have subsequently been developed to enhance segmentation and classification capabilities, showing state-of-the-art performance in domains ranging from autonomous vehicle perception to detailed urban modeling.

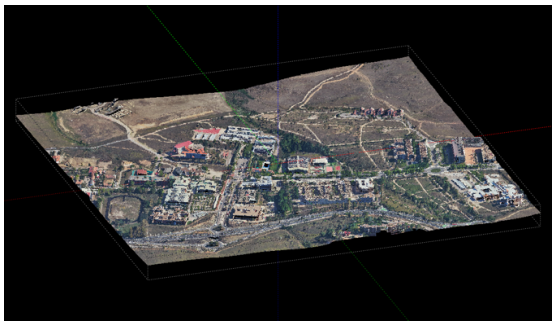


Fig. 2. Tools such as Laslook [9] allow for easy data visualization, as well as fast data transformations between file formats.

Despite remarkable developments, the literature reveals that a universally accepted methodological framework for deploying LiDAR in varied applications is still emerging. Current research continues to explore strategies for controlling precision and enhancing classification accuracy through comparative studies and real-world field validations.

III. Methodology

In response to the pressing need for a standardized framework in LiDAR data processing, current research endeavors are focusing on evaluating various processing applications to control accuracy levels effectively.

To achieve this, the study is structured in sequential stages. Initially, a selection of LiDAR datasets will be obtained from a pre-defined, geographically controlled environment, such as the University of Extremadura's campus in Cáceres. These will include raw point clouds that reflect diverse land cover types and elevation profiles.

Following acquisition, a pre-processing phase will be conducted to normalize the data. This includes noise reduction, ground filtering, and the segmentation of point clusters based on spatial density, return intensity, and elevation values (z-axis). Subsequently, multiple software applications will be employed to assess and compare the accuracy of height measurements.

Subsequently, once the data is adequately structured, metadata extraction will be carried out, where point classification algorithms will be applied to assess their reliability and efficiency when differentiating between terrain elements.

Following classification, a series of deployment models will be designed and tested using real-world examples in order to evaluate the operational viability of the proposed methodological steps under practical conditions. The research will ultimately conclude with the formalization of a methodological guide that outlines each procedural step mentioned.

IV. Discussion and Conclusions

This project proposes a comparison between LiDAR data samples and other datasets obtained through traditional methods, aiming to assess whether the classification accuracy is sufficient for practical application. A future objective is to integrate Artificial Intelligence algorithms, including cybersecurity, incorporate features to enhance the handling of these types of datasets.

Acknowledgments

This initiative is carried out within the framework of the funds of the Recovery, Transformation and Resilience Plan, financed by the European Union (Next Generation) - National Cybersecurity Institute (INCIBE) in the project C107/23 “Artificial Intelligence Applied to Cybersecurity in Critical Water and Sanitation Infrastructures”.

References

[1] MatLab. Lidar data processing overview. https://www.mathworks.com/help/lidar/ug/lidar-processing-overview.html. Accessed: 2025-04-15.

[2] Datumate. Digital surface model (DSM) and digital terrain model (DTM). https://www.datumate.com/glossary/digital-surface-model-dsm/. Accessed: 2025-04-15..

[3] Blickfeld. Lidar for security applications. https://www.blickfeld.com/applications/security/. Accessed: 2025-04-15.

[4] Reuters. Chinese lidar sensors pose hacking risk to U.S. defense equipment – report. https://www.reuters.com/world/china/chinese-lidar-sensors-pose-hacking-risk-us-defense-equipment-report-says-2024-12-02/, December 2024. Accessed: 2025-04-15.

[5] SurfaceModeling.com. Tin (triangulated irregular network) guide. http://www.surfacemodeling.com/Help/Guide/start.htm. Accessed: 2025-04-15.

[6] Esri. What is a las dataset? https://desktop.arcgis.com/es/arcmap/latest/manage-data/las-dataset/what-is-a-las-dataset-.htm. Accessed: 2025-04-15.

[7] Terrasolid. Terrascan product page. https://terrasolid.com/products/terrascan/. Accessed: 2025-04-15.

[8] J. Heinzel and S. Hese. Using supervised and unsupervised machine learning to classify lidar measurements. Atmospheric Measurement Techniques, 14:391–406, 2021. Accessed: 2025-04-15.

[9] Rapidlasso GmbH. laslook – visualizing las and laz files. https://rapidlasso.de/laslook/. Accessed: 2025-04-15.

Challenges of Artificial Intelligence in cybersecurity

David Marques
Coimbra Business School | ISCAC,
Polytechnic of Coimbra,
Coimbra, Portugal
dmrmarques@gmail.com

Sofia Félix
Coimbra Business School | ISCAC,
Polytechnic of Coimbra,
Coimbra, Portugal
felixsofia@sapo.pt

Georgina Morais
Coimbra Business School | ISCAC,
Polytechnic of Coimbra,
Coimbra, Portugal
mmorais@iscac.pt

Abstract — **Effective responses to daily information security demands are an increasingly present need, and must be ensured by mechanisms that meet the expectations of stakeholders. Certifications such as digital maturity and cybersecurity seals are examples of these mechanisms, allowing organizations to mitigate various physical and digital risks to which they are exposed. In this context, this research aims to highlight the challenges that Artificial Intelligence (AI) poses to cybersecurity.**

The recent publication of ISO 42001:2023 — the first global standard dedicated to the management of AI systems in organizations — represents a step forward in the standardization of AI governance practices. This standard promotes the responsible development and use of AI, also addressing ethical issues and emerging challenges. In short, information security must be guaranteed in a manner that is aligned with the needs and expectations of stakeholders, with quality being a decisive factor in the certification of systems, such as AI, digital maturity and cybersecurity.

Keywords - cybersecurity, ethics, artificial intelligence.

I – Introduction

The increasing digitalization of society has profoundly transformed organizational structures, requiring continuous adaptation to information and communication technologies. In this context, AI emerges as a disruptive technology with promising applications across various domains, including cybersecurity (CS). However, its integration also introduces new risks, underscoring its pivotal role in information security.

AI enhances real-time threat detection, predictive analysis of malicious behavior, and automated incident response. However, it also raises significant challenges, such as vulnerability to adversarial attacks, ethical concerns regarding the use of sensitive data, and the potential for malicious use of AI itself—for instance, through deepfakes or adaptive malware. Consequently, a balanced approach is required, one that fosters technological innovation while safeguarding fundamental rights, with a focus on algorithmic transparency, human oversight, and responsible governance. The intersection between AI and CS thus demands critical and multidisciplinary reflection. In this regard, the present study aims to analyze the main challenges arising from the application of AI to CS, structured into five sections: (I) Introduction; (II) Advantages and Challenges of AI in CS; (III) Regulatory Framework and Legislation; (IV) Portuguese Companies and the Global AI Context – AI Adoption; and (V) Conclusions.

II – Advantages and Challenges Of AI in CS

The integration of AI into CS offers vast potential for enhancing system and network defenses, but it also introduces a range of challenges, both technical and ethical in nature.

In this section, we will explore the main advantages and risks associated with the use of AI in protecting against cyber threats. Key advantages of AI include: rapid processing of large volumes of data; detection of anomalies and unusual activities; automation of threat response; and real-time insights into security events

Despite its potential, AI in CS faces several obstacles, the complexity of which is amplified by the speed at which digital threats evolve. Key challenges include: privacy and data protection; transparency; algorithmic manipulation (overreliance on automation); and resilience to evolving threats. Risks associated with AI include not only vulnerabilities intrinsic to algorithms, but also ethical and social issues, such as the impact of automation on society, consent management, and the increasing sophistication of cyber threats driven by AI itself.

Indeed, although AI contributes to strengthening security systems, it also intensifies the complexity of cyber attacks, and there is a need for greater transparency in algorithms, specific regulation and cooperation between different sectors to face the challenges of the digital age and promote a safer environment [1].

Additionally, ethical and privacy concerns must be taken into account, with the most relevant issues relating to: data privacy and protection; accountability and responsibility in the use of AI technologies; and legal considerations. Thus, the implementation of AI algorithms in security systems presents both ethical and technical challenges. Other key issues include: the evolving nature of cyber threats; the increasing complexity of IT environments; and the limitations of traditional rule-based security systems. Threats inherent to the use of AI include: vulnerabilities within AI systems themselves; overreliance on automation; privacy and consent issues; bias manipulation; lack of transparency and explainability; social and ethical impact; and the rise of advanced threats.

The new ethical challenges associated with AI derive largely from the fact that AI algorithms are increasingly being applied

to tasks involving social and cognitive dimensions that, until now, were performed exclusively by humans. In these contexts, algorithms begin to incorporate social requirements, making it essential to understand the impact of the use of AI both in organizations and in people's daily lives. In particular, it is crucial to identify the ethical principles that should govern these applications, as well as to establish effective mechanisms for their monitoring and intervention. Therefore, it is pertinent to carry out in-depth studies, at individual and organizational level, that analyse the effects of AI systems, contributing to a responsible and ethical use of these technologies [2].

Although AI offers a number of technical benefits that can improve the effectiveness and efficiency of CS operations, namely: in real-time threat detection and prevention; in predictive analysis; in automatic incident response. Despite the advantages, the implementation of AI in CS is not free from significant challenges, such as: vulnerability to adversarial attacks; complexity and cost of implementation; lack of transparency and explainability; ethical and privacy challenges. On the other hand, another important challenge to be considered is the impact of automation on the CS job market. With the increasing adoption of AI, certain functions, such as system monitoring and initial analysis of security alerts, can be automated, leading to a change in the skills required of professionals in the area. This may result in the need for worker retraining and school adaptation to prepare professionals to deal with AI technologies effectively [3].

Although the advantages of using AI in CS are widely acknowledged, several challenges are associated with its implementation, including: complexity of integration; lack of high-quality data; risks of false positives and false negatives; and adversarial AI. Among the key challenges, the European AI regulation and the NIS2 Directive also stand out as critical regulatory considerations.

III – Regulatory Framework and Regulation

With the increasing use of AI in CS, regulatory and regulatory issues arise that need to be addressed to ensure compliance with data protection and privacy laws, such as the regulation of AI algorithms; data protection and privacy; and legal liability in the event of AI failures.

Regarding the regulation of AI algorithms, it is important to note that AI can have a significant impact on security-related decision-making. Regulations guiding the use and development of AI algorithms are essential to ensure that AI systems operate fairly and transparently [4]. In terms of data protection and privacy, the collection and processing of large volumes of data for training AI models raise serious privacy concerns. Laws such as the European Union's General Data Protection Regulation (GDPR) require organizations to implement strict data protection practices. AI-based solutions must ensure that personal data is handled in compliance with privacy regulations to avoid potential rights violations [5]. Concerning legal liability in the event of AI failures, it remains a considerable challenge to determine responsibility when AI-based systems fail. If an AI system fails to detect a cyber threat, or makes incorrect decisions resulting in harm, a clear line of accountability must be established [6].

At the regulatory level, the European Union (EU) has approved the AI Regulation, which is a key element of EU policies aimed at promoting the development and adoption, across the single market, of safe and lawful AI that respects fundamental rights, and is intended to harmonise rules on AI [7]. This legislation follows a "risk-based" approach and aims to promote the development and adoption of safe and trustworthy AI systems, both by public administrations and private individuals.

Regarding the normative part, the NIS2 directive stands out in the European context, which establishes CS measures in the Member States. Additionally, the international standard ISO/IEC 42001:2023, which establishes requirements for AI management systems, reinforces the importance of integrating quality, information security and algorithmic governance, promoting a holistic approach to technological risk management.

Regarding AI ethics, UNESCO recommends eleven areas of policy action, namely: ethical impact assessment; governance and ethical management; data policy; development and international cooperation; environment and ecosystems; gender; culture; education and research; communication and information; economy and work; health and social well-being. In terms of CS, it stipulates the fundamental role that it plays in ensuring that AI systems are resistant to the actions of malicious third parties who attempt to exploit the vulnerabilities of systems with the aim of altering their use, behavior and performance or compromising their security properties. CS involves the development of actions, good practices, concepts, guidelines, tools, policies or technologies to be implemented by organizations, with a view to preventing potential cyberattacks, especially techniques and behaviors applied by individuals in terms of their digital security [8]. On the other hand, the increase in cyberattacks has been observed, and the positioning of risk management in relation to information security should be assessed in organizational terms.

Unlike traditional CS approaches, which often rely on manual intervention and predefined rule sets, AI ushers in a new era of automation and intelligence-driven defense mechanisms. At the heart of this transformation are techniques such as Machine Learning (ML) and Deep Learning (DL), which empower AI systems to analyze vast amounts of data at unprecedented speeds and complexity. Cyberattacks against AI systems can leverage specific AI assets, such as training datasets (e.g., data contamination) or trained models (e.g., adversarial attacks), or exploit vulnerabilities in the AI system's digital assets or the underlying information and communications technology infrastructure. In order to ensure a risk-appropriate CS level, vendors of high-risk AI systems must take appropriate

measures, while also taking into account the underlying ICT infrastructure. In turn, AI can be used in CS to detect and mitigate threats, build resilience into organizational operations, and accelerate processes such as phishing detection through image recognition and automated tracking, such as: cyber threat intelligence; anomaly detection; vulnerability management; intrusion prevention systems; phishing protection; malware analysis; automated incident response; network security; ML for fraud detection; future threat prediction; advanced biometric authentication; behavioral analytics for endpoint security; user behavior analysis; supply chain attacks; security training and awareness; cloud security; password management; spam filtering.

In the context of the European Union, CS is addressed through the NIS2 Directive, the second iteration of the Network and Information Systems Directive—a landmark piece of cybersecurity legislation aimed at establishing a higher level of cyber resilience across organizations throughout the EU. Under NIS2, additional sectors have been included: wastewater management; ICT service management; public administration; space; postal and courier services; waste management; production, manufacturing, and distribution of chemicals; production, processing, and distribution of food products; manufacturing industry; digital service providers; and research.

Cybersecurity also benefits significantly from the use of AI, as it enables: advanced threat detection; automated incident response; behavioral analysis for security; adaptive defense mechanisms; and countermeasures against malicious AI.

The NIS2 Directive reinforces some of the measures already provided for in Decree-Law No. 65/2021 [9], namely in relation to risk analysis and incident handling, including a minimum set of topics that must be issued by organizations. Thus, based on the provisions of the NIS2 Directive, entities must, at a minimum, issue the following topics: risk and security analysis; incident handling; business continuity; supply chain security; security in acquisition, development and maintenance; assessment of the effectiveness of CS risk management measures; basic cyber hygiene practices and CS training; cryptography and, where applicable, encryption; human resource security; use of multi-factor authentication or authentication solutions [10].

The quality of the information that institutions hold and make available must be measured. In this sense, organizations have at their disposal certain instruments, such as certain certifications (e.g.: ISO 9001; ISO 27001; ISO 42001; among others), which allow them to guarantee security and trust to their stakeholders. These certifications allow entities to mitigate many of the physical and digital risks to which they are exposed, contributing to the increase in information security and its protection, increasing the recognition of their commitment to information security and the trust of stakeholders in their activities and CS. On the other hand, the international publication of ISO 42001 standardizes AI management practices, as it standardizes its responsible development and use, enabling certification. This standard addresses the challenges posed by AI, including ethical issues, increasing trust in its activities, as well as greater transparency in the implementation of AI.

ISO/IEC 42001 is an international standard that specifies requirements for establishing, implementing, maintaining, and continually improving an Artificial Intelligence Management System (AIMS) [11] within organizations. AIMS is designed for entities that provide or use AI-based products or services, ensuring the responsible development and use of AI systems. ISO 42001 is applicable to organizations of any size, across industries, public sector bodies, businesses, and non-profit institutions.

The objective of ISO/IEC 42001 is to provide organizations with comprehensive guidance on the responsible and effective use of AI, even in contexts where technology is rapidly evolving. The implementation of ISO/IEC 42001 ensures ethical and responsible use of AI, promotes greater trust in AI applications, supports regulatory compliance, enables effective risk management, and encourages innovation. Moreover, the standard places strong emphasis on the implementation of robust security measures to protect AI systems from unauthorized access, data breaches, and other cyber threats.

Amid this digital revolution lies a critical weakness characterized by the pervasive threat of cyberattacks [12]. Traditionally, cybersecurity has relied heavily on signature-based and rule-based detection methods to identify threats. However, as attackers become more sophisticated and threats continue to evolve, these approaches are revealing their limitations [13]. Signature-based detection struggles with novel attacks, while rule-based methods often generate false positives and fail to detect complex threats.

ISO/IEC 42001, on the other hand, is a management system standard. Implementing this standard means implementing policies and procedures for an organization's good governance in relation to AI, using the Plan-Do-Check-Act methodology, and practically managing risks and opportunities related to AI across the organization, providing value to any organization by mitigating many of the physical and digital risks. Thus, it contributes to increasing information security and protecting organizations.

IV – Portuguese Companies and The Global Context – AI Adoption

The digital presence of Portuguese companies continues to grow at a rapid pace, driven by the implementation of innovative digital strategies, such as AI. According to the latest study “Digital Economy in 2024”, published by the Digital Economy Association in Portugal (ACEPI), 17% of Portuguese companies have already adopted AI practices in their operations, an indicator that represents more than double the European average [14]. In addition to increasing

competitiveness and resilience in the market, AI has a direct impact on online commerce.

Another study, "Unlocking Portugal's Ambitions on Artificial Intelligence (AI) in the Digital Decade by 2025", by Amazon Web Services (AWS) and Strand Partners, states that in 2024, 96,000 Portuguese companies adopted AI for the first time, which corresponds to 41% of all companies in the country [15] and a growth of 17% compared to 2023. According to the report, 94% of companies that incorporate AI into their activity report an average increase in revenue of around 30%, with 77% registering significant improvements in productivity. The study reveals that 62% of business organizations are startups and that 65% of small and medium-sized enterprises (SMEs) are still at basic levels of AI adoption. According to the analysis, investment in digital technology grew by 61% in the last year, a difference of ten percentage points compared to the European average, which does not exceed 51%. "The rapid pace of adoption of digital technologies, especially AI, could unlock €61 billion for the Portuguese economy if it continues at this pace," the report says. However, AI adoption in Portugal is not occurring at the same rate across the country. According to the study, only 11% of large companies are using technology in a transformative way and only 19% have a global AI strategy. The lack of digital skills among employees is cited by 42% of companies as one of the barriers to implementation. Around 71% of Portuguese employers admit to having difficulty in hiring professionals with the right skills and retaining them in their companies. The European average is 44%. Initial investment is another obstacle to AI adoption, with 34% of companies citing costs as a barrier, while 21% cite a lack of clarity about the return on investment as an obstacle. Regulatory uncertainty also deters more than a third of the 1,000 national and 1,000 Portuguese companies interviewed for the study, with 37% of business organizations hesitating to implement it because they do not know what they might expect. In the global context and according to the Global Insights Whitepaper [16] study - Building a People-Centric Strategy for AI-Driven Productivity, by Experis, large companies (between 1,000 and 5,000 employees) are those that most use AI, with more than half (54%) stating that they are already using it currently. In companies with less than 50 employees (micro and small businesses), this adoption is 10 percentage points lower.

The Experis study analyzed responses from over 40,000 employers across 42 countries to gain clearer insights into the current and future adoption of AI across various sectors. Respondents revealed that 65% of the workforce, across all hierarchical levels, believe AI will have a positive impact on the future of work [17]. In Portugal, senior and middle management are the most optimistic groups (70%), followed by administrative professionals (68%). Factory and frontline workers are more cautious, with only 53% expressing the same optimism.

Regarding challenges, privacy and regulation are identified as the main barriers to AI adoption by employers (35%). Investment costs (30%) and employee resistance to change (30%) are also cited as challenges. This scenario mirrors global trends, where costs are the primary concern for one-third of employers, followed by privacy and regulatory issues. According to Experis, and contrary to the common perception that AI-based technologies will lead to job losses, 48% of employers in Portugal believe AI will lead to growth in their teams. Only 25% anticipate a reduction, and 23% expect no impact.

Regarding its application, AI proves to be a transversal technology with the potential to profoundly transform various sectors of society, including healthcare, transportation, finance, education, agriculture, and industry.

• Healthcare: AI can contribute to the early diagnosis of diseases, development of personalized treatments, efficient management of healthcare services, and analysis of large volumes of clinical data, promoting more precise and predictive medicine.

• Transportation: in the transportation sector, AI enables improved safety and efficiency of autonomous vehicles, optimization of routes and logistics operations, and holds the potential to profoundly transform urban mobility systems and public transportation.

• Finance: AI is widely used in fraud detection, risk management, predictive investment analysis, personalization of financial services, and automation of banking processes, fostering greater agility and security in operations.

• Education: in education, AI enables personalized teaching and learning, adaptive tutoring, identification of specific student difficulties, and improvement of assessment and pedagogical management systems.

• Agriculture: AI applications in agriculture allow optimization of resource use, monitoring of crop growth, more accurate yield predictions, and early detection of pests and diseases, contributing to more sustainable and efficient farming.

• Industry: in the industrial sector, AI plays a key role in automating production processes, optimizing supply chains, predicting equipment failures, and improving energy efficiency, also promoting more sustainable practices.

However, despite the numerous opportunities, the application of AI also entails significant challenges in these sectors, which must be carefully considered to ensure ethical, safe and effective implementation, such as:

• Healthcare: One of the main challenges is protecting patients' privacy and sensitive data, especially in the context of sharing and processing large volumes of medical information. In addition, there is a need to ensure the transparency of algorithms used in clinical decisions, as well as rigorous scientific validation of AI systems before their practical application.

• Transportation: the use of AI in autonomous vehicles raises concerns related to legal liability in the event of accidents, the cybersecurity of integrated systems and the need for adequate smart infrastructure. Social acceptance and user trust are also critical factors for its adoption.

• Finance: in this sector, the challenges involve mitigating risks associated with the opacity of decision algorithms (the so-called black box problem), preventing algorithmic discrimination and complying with regulations such as the GDPR. Cybersecurity remains a central concern, given the sensitivity and value of financial data.

• Education: AI can exacerbate existing inequalities if equitable access to technologies and adaptation of systems to the needs of all students are not ensured. In addition, there are concerns about the dehumanization of education and the excessive replacement of human interaction with automated tools.

• Agriculture: the adoption of AI-based solutions may be hampered by the lack of digital infrastructure in rural areas, as well as the need for technical training on the part of farmers. The reliance on high-quality data to train effective models is also a major obstacle.

• Industry: the integration of AI into supply chains requires significant investments in technological modernization and reskilling of the workforce. In addition, intensive automation raises questions about the future of employment, the redistribution of work roles, and the security of connected industrial systems.

Regarding its use, AI is used by Portuguese companies. These are part of economic groups, namely, for example: NOS; EDP; Jerónimo Martins. NOS uses AI to automate customer service. Regarding EDP, it uses AI to predict failures in the electricity grid. Jerónimo Martins uses AI to optimize logistics. In terms of effectiveness, for AI to be used effectively by Portuguese companies, it is important to:
• invest in training, that is, organizations must invest in training so that their employees can use AI effectively;
• create a culture of innovation: organizations must create a culture of innovation that encourages experimentation and adoption of new technologies;
• collaborate with universities and research centers: organizations must collaborate with higher education to develop and apply new AI technologies.

Therefore, the use of AI in Portuguese companies can boost innovation and efficiency, increase growth and competitiveness, and prepare for the future of work. He adds that AI can be especially useful for Portuguese companies by helping to:
• overcome the challenges of a small market: AI can help Portuguese companies reach a global audience and compete with companies from larger countries;
• reduce costs: AI can help Portuguese companies reduce production and operating costs;
• increase productivity: AI can help Portuguese companies increase workforce productivity.

In conclusion, AI is a powerful tool that can help Portuguese companies grow and thrive. By investing in AI and using it effectively, companies can prepare for the future of work and become more competitive in the global market. So, one might ask the following question: is AI an ally or an obstacle for companies?

AI can be both an ally and an obstacle for businesses, depending on how it is implemented and managed. As an ally, AI can automate repetitive tasks, increase operational efficiency, improve data-driven decision-making, and even drive innovation. However, if not implemented properly, AI can pose challenges such as high implementation costs, data privacy and security concerns, over-reliance on technology, and even replacing human jobs in certain areas. Therefore, the success of AI in companies depends on a solid strategy, adequate investment, ethics in its use, and a clear understanding of the benefits and challenges involved.

In terms of CS and its connection with AI, the analysis of emerging threats and defensive strategies, it is observed that, although AI contributes significantly to the improvement of security systems, it also increases the sophistication and complexity of cyberattacks. The study highlights the urgency of greater transparency in algorithms, the creation of specific regulations and the strengthening of cooperation between the public, private and academic sectors. These measures are essential to face the challenges of the digital age and promote a safer and more resilient cyber ecosystem.

VI – CONCLUSIONS

This article has explored the intersection of AI and cybersecurity, highlighting both the opportunities and challenges involved in using AI to strengthen the protection of systems, networks, and data against emerging digital threats.

The application of AI in cybersecurity opens up a wide range of possibilities for strengthening organizational resilience, especially with regard to early incident detection, automated response, and proactive vulnerability management. However, there are still technical, ethical, and legal challenges that need to be addressed to ensure the safe and effective adoption of these technologies. The future of cybersecurity will depend directly on the continued evolution of AI, both in terms of algorithmic sophistication and in terms of creating regulatory frameworks that ensure the responsible and transparent use of these tools.

As AI advances, it is imperative that organizations stay up to date with technological and regulatory trends, while simultaneously fostering a culture of security that respects privacy and regulatory compliance. The increasing

incorporation of AI in cybersecurity must be accompanied by a proactive approach to resolving legal and ethical issues, ensuring public trust and the integrity of digital systems. Institutions therefore have a responsibility to ensure the security of the information they produce and hold, which reinforces the centrality of cybersecurity in the current digital paradigm driven by AI.

In this context, the European Regulation on Artificial Intelligence [18] stands out, the first global legislation dedicated exclusively to regulating this technology. This regulatory framework has the potential to set a global standard, promoting the harmonization of international legal standards and ensuring that AI is developed and used in an ethical, safe and trustworthy manner [7].

The adoption of AI by organizations also offers new possibilities for preventive risk monitoring and for developing employees' cybersecurity skills. With its ability to process large volumes of data and identify complex patterns, AI can profoundly transform the way in which digital threats are addressed. However, it is crucial to mitigate the risks inherent in the use of automated systems, such as malfunctions, opaque decision-making processes and unforeseen vulnerabilities.

Clear and effective regulations are not only essential to safeguard security and privacy, but also to foster public trust in the use of AI in the field of computer science. In short, regulating AI in this field represents an essential strategic step to ensure that the benefits of these technologies can be fully exploited, while minimizing their risks and strengthening the protection of society against future digital threats.

In the business context, the application of AI represents a significant opportunity, but also raises complex technical, ethical and regulatory challenges. An analysis of references such as ISO/IEC 42001 and the NIS 2 Guideline highlights the growing need for a robust regulatory framework that balances innovation with responsibility. However, throughout this study, a gap was identified in the national scientific literature on the ethical and regulated application of AI in the business sector, making it difficult to systematize good practices and anticipate emerging risks. To overcome this limitation, it is crucial to encourage interdisciplinary scientific production, as well as collaboration between universities, companies and regulatory bodies. Such an effort could generate practical and up-to-date knowledge that supports the safe and ethical adoption of AI. In this sense, the importance of an integrated approach that unites research, regulation and responsible business application is reinforced.

BIBLIOGRAPHICAL REFERENCES

[1] Cruz, J.; Casemiro, J.; Gallizzi; J.; Kalili; R. (2024), Inteligência artificial e cibersegurança: análise de ameaças emergentes e estratégias defensivas, Revista DELOS, Curitiba, v.17, n.61, p. 01-16

[2] Piteira, M. , Aparicio, M. e Costa, C. J. (2019), A ética na inteligência artificial: Desafios, CISTI'2019 - 14ª Conferência Ibérica de Sistemas e Tecnologias de Informação, Junho 2019, Coimbra, Portugal

[3] N. A. Collins and S. K. Brown, "The Impact of AI Automation on Cybersecurity Job Markets," IEEE Computer Society Magazine, vol. 40, no. 7, pp. 87-95, Jul. 2025.

[4] L. F. Gonzalez and P. R. Diaz, "Regulation of AI in the European Union: The AI Act and Its Implications for Cybersecurity," European Journal of Law and Technology, vol. 29, no. 4, pp. 155-167, Oct. 2023.

[5] A. J. Clark and J. F. Dunn, "Privacy and Data Protection in AI-Powered Cybersecurity Systems," IEEE Transactions on Privacy and Security, vol. 21, no. 7, pp. 223-235, Jul. 2024.

[6] M. S. Patel, "Legal Liability in AI-Driven Cybersecurity Systems: A Review of Current Challenges," International Journal of Cybersecurity Law, vol. 8, no. 2, pp. 54-63, Apr. 2023.

[7] Conselho da União Europeia (2024). Regulamento Inteligência Artificial da UE. Disponível em: https://www.consilium.europa.eu/pt/policies/artificial-intelligence/#0.

[8] Veale, M.; & Brown, I. (2020). Cybersecurity. Internet Policy Review. Journal on Internet regulation. https://doi.org/10.14763/2020.4.1533.

[9] Decreto-Lei n.º 65/2021, Presidência do Conselho de Ministros, de 30 de julho.

[10] Diretiva (UE) 2022/2555, do Parlamento Europeu e do Conselho, de 14 de dezembro de 2022, Diretiva SRI 2 (NIS 2).

[11] ISO/IEC 42001: 2023 (2023). Information technology — Artificial intelligence — Management system. Disponível em: https://www.iso.org/standard/81230.html.

[12] Urbinati, A.; Chiaroni, D.; Chiesa, V. and F. Frattini (2018). The Role of Digital Technologies in Open Innovation Processes: An Exploratory Multiple Case Study Analysis, R D Manag, doi: 10.1111/radm.12313.

[13] Howard, D. J. (2018). Development of the Cybersecurity Attitudes Scale and Modeling Cybersecurity Behavior and its Antecedents. University of South Florida, 86.

[14] Digital em Portugal, "Digital em Portugal – Transformação Digital da Economia e da Sociedade," [Online]. Available: https://digitalemportugal.pt/.

[15] Meios & Publicidade, "41% das empresas portuguesas já usam inteligência artificial," 14-May-2025. [Online]. Available: https://www.meiosepublicidade.pt/2025/05/14/41-das-empresas-portuguesas-ja-usam-inteligencia-artificial.

[16] HR Portugal, "Organizações em Portugal mais cautelosas com a adopção da IA do que média global," [Online]. Available: https://hrportugal.sapo.pt/organizacoes-em-portugal-mais-cautelosas-com-a-adopcao-da-ia-do-que-media-global/

[17] AICEP Portugal Global, "Utilização de IA pelas empresas", Portugal Global, 3 jul. 2024. [Online]. Disponível em: https://www.portugalglobal.pt/pt/noticias/2024/julho/utilizacao-de-ia-pelas-empresas/

[18] European Commission, "Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts," COM(2021) 206 final, Apr. 2021. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2021%3A206%3AFIN

Session 3

Digital Wallets in the Metaverse: A Blockchain-Based Approach to Enhance Payment Systems

Mouloud AFOULOUS
Department of Information Science and Technology,
Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR,
Lisboa, Portugal
mouloud_afoulous@iscte-iul.pt

Catarina Ferreira da Silva
Department of Information Science and Technology,
Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR,
Lisboa, Portugal
catarina.ferreira.silva@iscte-iul.pt

Abstract— The metaverse, an expansive virtual realm, intertwines seamlessly with blockchain technology and digital wallets, revolutionizing the landscape of digital interactions. This research navigates the convergence of these domains - metaverse, blockchain technology and digital wallets -, particularly emphasizing the imperative to fortify the interoperability and efficiency of payment systems within this evolving landscape.

In this dynamic metaverse, users immerse themselves in virtual experiences, transcending the boundaries of the physical world. Simultaneously, blockchain technology stands out for its decentralized architecture and cryptographic foundations, offering a transformative influence across industries. Digital wallets, as crucial components in this paradigm, facilitate seamless transactions and asset management.

This study aims to decipher the intricate relationship between the metaverse, blockchain, and digital wallets. By delving into existing literature, it identifies nuanced gaps and challenges in current payment systems within virtual realms. The research sets forth a novel blockchain-based approach to fortify the security of digital wallets within the metaverse. The emphasis lies in integrating decentralized technologies to instil trust, transparency, and efficiency in virtual transactions.

A meticulously designed project execution plan charts the course for the research journey over four years. The plan encapsulates a comprehensive exploration, including literature reviews, prototype design and development, usability studies, security assessments, and regulatory analyses.

This endeavour aspires to contribute not only to a nuanced comprehension of the metaverse but also to offer innovative solutions for interoperable digital transactions. By developing a tangible, secure digital wallet prototype, the research aims to influence and elevate the landscape of digital transactions within virtual environments.

Our commitment is to bridge the theoretical realms of the metaverse, blockchain, and digital wallets with practical solutions. This research seeks to echo in the academic sphere and beyond, fostering a more secure, transparent, and efficient future for financial interactions within the metaverse.

I. Introduction

The motivation behind this research stems from the escalating significance of secure digital transactions within the expansive landscape of the metaverse and its profound implications for the banking and financial sectors [25]. As our world becomes increasingly intertwined with virtual realms, the metaverse emerges as a dynamic and immersive space where users engage in experiences that transcend traditional physical boundaries.

In this evolving digital ecosystem, the demand for secure and efficient financial transactions within the metaverse has become more critical than ever. Existing payment systems, while robust in conventional settings, face unique challenges when applied to decentralized virtual environments. A clear need exists for a secure, transparent, and seamless payment infrastructure tailored to the specific requirements of the metaverse, especially as digital assets and peer-to-peer interactions become central to its economy.

Traditional digital wallets often remain vulnerable to sophisticated cyber threats and lack transparency and interoperability. Conventional banking structures are ill-equipped to handle the fluid and decentralized nature of virtual transactions, raising critical concerns around privacy, identity, and user trust. This research aims to address these challenges by proposing an innovative, blockchain-based solution that integrates decentralized technologies into digital wallets to secure financial interactions within the metaverse.

Blockchain technology, with its decentralized architecture and cryptographic foundations, offers a compelling opportunity to enhance trust and efficiency in virtual economies. By embedding blockchain capabilities within digital wallets, we aim to create a resilient and transparent payment ecosystem, offering benefits such as tamper-proof transactions, smart contract automation, and decentralized identity management.

For enterprises and developers navigating this emerging space, the proposed framework not only mitigates existing security concerns but also offers strategic advantages in adapting to the metaverse's financial paradigm. A secure and interoperable wallet infrastructure may catalyze broader adoption and innovation within the virtual economy.

To better orient the reader, the structure of this paper is as follows: **Section II** delineates the existing literature in the fields of digital wallets, blockchain, and the metaverse. **Section III** identifies the research objectives underpinning

this study. **Section IV** provides our research proposal, delving into the key concepts and theoretical underpinnings. In **Section V**, we outline the task execution plan across the four-year research period. The paper concludes with reflections on the expected contributions and suggestions for future research directions.

II. STATE OF ART

Understanding the intersection of digital wallets, blockchain technology, and the metaverse is crucial to designing secure and efficient financial systems for virtual environments. This section surveys the foundational literature and recent advancements in each domain, identifies key challenges, and highlights emerging technological solutions.

A. Digital Wallets

Digital wallets, or **e-wallets**, have become essential in modern financial systems by enabling users to store, manage, and transact with digital currencies or tokenized assets [1][2] [25]. They facilitate everyday payments, peer-to-peer transfers, ticketing, and online purchases through mobile and web platforms.

Major industry players such as **Apple Pay, Google Pay, Alipay, and WeChat Pay** have popularized digital wallets, promoting seamless transactions and convenience. Advanced security features such as biometric authentication, tokenization, and multifactor authentication have improved wallet safety; however, centralized infrastructures continue to pose risks of data breaches and unauthorized access [3][4].

Recent studies show increased user adoption of wallets in metaverse-related contexts, with a growing need for integration with identity management and cross-platform operability [5][6]. Furthermore, recent work in financial UX emphasizes trust design, user-centered cryptographic workflows, and embedded compliance as critical for mass adoption in immersive platforms [7].

Despite their widespread use, traditional e-wallets are generally not designed for interoperability with decentralized systems or for integration into immersive virtual environments, which limits their applicability in the metaverse.

B. Blockchain Technology

Blockchain serves as a tamper-resistant, decentralized ledger ideal for enhancing trust and transparency in financial transactions [8]. Originally designed to support cryptocurrencies like Bitcoin, blockchain is now integral to a broader range of applications, from decentralized finance (DeFi) to tokenized identity and asset systems [9][10].

Smart contracts—self-executing programs stored on blockchains—offer the ability to automate agreements and transactions without intermediaries [11]. In metaverse environments, this enables programmable asset ownership, microtransactions, and event-triggered payments.

Emerging blockchain standards such as **ERC-4337** for account abstraction and **Decentralized Identifiers (DIDs)** defined by W3C are paving the way for secure user interactions within virtual ecosystems [12]. However, scalability, energy consumption, and on-chain/off-chain interoperability remain open technical challenges [13][14].

C. The Metaverse

The metaverse is defined as a persistent, interconnected virtual space, composed of virtual reality (VR), augmented reality (AR), and mixed reality (MR) environments [15]. It supports rich social and economic interactions, powered by digital identities, avatars, and virtual currencies.

Recent analyses confirm the trend toward mixed-reality commerce, where digital wallets must support not only traditional finance operations but also avatar-linked transactions, game token economies, and NFT ownership mechanisms [16][17].

As user-generated content and digital asset ownership become more prevalent, the need for secure and decentralized transaction mechanisms within the metaverse becomes evident [18].

D. Challenges in Metaverse Payment Systems

A recurring theme in literature is the inadequacy of traditional financial infrastructure when applied to the metaverse. Centralized platforms are vulnerable to cyberattacks, fraud, and regulatory opacity [19]. Furthermore, the lack of cross-platform compatibility hinders fluid economic interaction across virtual environments.
Key issues include:

- **Security vulnerabilities** in legacy wallet architecture.
- **Lack of standardized protocols** for inter-world commerce.
- **Uncertain legal frameworks** around token-based payments.

This calls for collaborative efforts between academia, industry, and regulators to ensure scalable, compliant, and user-centric payment infrastructure.

E. Emerging Solutions and Research Gaps

Recent developments suggest that the integration of blockchain-enabled wallets with metaverse platforms is not only feasible but essential for secure and interoperable financial interactions [20]. Decentralized identity (DID) frameworks such as Verifiable Credentials (W3C) and soulbound tokens offer potential for securing avatar-linked identities.

Efforts like Web3Auth, Worldcoin, and Self-Sovereign Identity (SSI) protocols explore how to embed identity, ownership, and compliance in a user-centric and decentralized way [21][22].

Additionally, ongoing work on metaverse-native compliance—including zero-knowledge proofs, on-chain

KYC, and programmable privacy—addresses the privacy-security duality in decentralized environments [23].

III. OBJECTIVES

The overarching objectives of this research proposal are to bridge existing gaps and overcome challenges in the realm of secure digital transactions within the metaverse, with a particular emphasis on the integration of secure digital wallets with blockchain and decentralized ledger technology.

1. Investigate Current Gaps and Challenges: The primary objective is to conduct an in-depth investigation into the existing gaps and challenges in the metaverse's payment systems. This involves a comprehensive analysis of literature pertaining to digital wallets, blockchain, decentralized ledger technologies and the metaverse, with the aim of identifying vulnerabilities, inefficiencies, and areas of friction within current payment infrastructures. By pinpointing these challenges, the research seeks to lay the groundwork for targeted interventions and advancements.

2. Propose a Blockchain-Based Solution for Metaverse Digital Wallets: Building upon the identified gaps, the research aims to propose a robust and innovative solution by integrating metaverse with blockchain-based digital wallets. This involves the exploration of decentralized ledger systems, cryptographic principles, and smart contract functionalities. The objective is to leverage the inherent security features of blockchain to fortify digital wallets within the metaverse. This proposed solution should not only enhance security but also contribute to the efficiency and transparency of financial transactions in virtual environments.

3. Enhance Security Measures: With a focus on security enhancement, the research aims to develop and implement advanced security measures within digital wallets. This includes the incorporation of biometric authentication methods, cryptographic protocols, and decentralized identity management. The objective is to create a resilient security framework that mitigates the risks associated with unauthorized access, fraudulent activities, and other security threats specific to metaverse transactions.

4. Explore Smart Contract Applications: Smart contracts, being integral to blockchain and decentralized ledger technologies, present a unique opportunity for automation and programmability within digital wallets. The research objectives include exploring the applications of smart contracts in enhancing the functionality of digital wallets. This involves developing use cases for self-executing agreements, automated payment processes, and conditional transactions to streamline and secure financial interactions in the metaverse.

5. Address Interoperability Challenges: Recognizing the importance of interoperability in virtual transactions, the research aims to address the challenges associated with the seamless flow of digital assets across different metaverse platforms. By proposing solutions and standards for interoperability, the objective is to create a unified metaverse financial infrastructure that facilitates cross-platform transactions, fostering a more cohesive and accessible virtual economy.

6. Evaluate and Validate the Proposed Solution: To ensure the practical viability and effectiveness of the proposed blockchain-based solution, the research objectives include the development of a prototype of a blockchain-based digital wallet within a metaverse. This prototype will undergo rigorous evaluation, incorporating usability studies, security assessments, and real-world simulation scenarios with metaverse platforms. The objective is to validate the solution's efficacy and identify areas for further refinement.

In summary, the research objectives are designed to holistically address the identified gaps and challenges in metaverse payment systems, culminating in the development and validation of a blockchain-based solution that enhances the security, efficiency, and interoperability of digital wallets within the dynamic landscape of the metaverse.

IV. RESEARCH PROJECT PROPOSAL

A. Basis for Work Development:

The theoretical foundation of this research rests upon the dynamic intersection of three pivotal concepts: secure digital wallets, blockchain technology, and the metaverse. To comprehensively understand and navigate this intricate landscape, it is imperative to define key terms and concepts that underpin this research.

Secure Digital Wallets: In the context of this research, secure digital wallets refer to virtual tools that enable users to store, manage, and transact digital assets securely, with a focus on the metaverse. Emphasis is placed on bolstering security measures, including biometric authentication, cryptographic protocols [1], and decentralized identity management, to mitigate vulnerabilities and ensure the integrity of virtual financial transactions.

Central to this research is the integration of blockchain technology into digital wallets. Blockchain, a decentralised and tamper-resistant ledger, serves as the backbone for enhancing the security and transparency of metaverse transactions. Key components include smart contracts, self-executing agreements that automate and secure transactions, and consensus mechanisms that validate and record transactions across the decentralised network.

Within the scope of this research, the metaverse is defined as an interconnected virtual space where users engage in immersive experiences, transcending the boundaries of the physical world. This includes (VR), (AR), and (MR) platforms that constitute persistent virtual worlds [8].

The conceptual framework emerges from the symbiotic relationship between these key elements. Secure digital wallets act as gateways to the metaverse, facilitating seamless and reliable transactions. Blockchain technology provides the decentralized infrastructure necessary to fortify the security and transparency of these transactions.

B. Work Description:

The proposed work seeks to address the identified gaps and challenges in metaverse payment systems and propose a reliable integration of blockchain-based digital wallets with metaverse. This entails a multifaceted approach that encompasses technological, cryptographic, and usability considerations.

Integration of Blockchain into Digital Wallets: The core focus of the research is on integrating blockchain technology into metaverse digital wallets to create a secure

and decentralized foundation for metaverse transactions. The integration involves the development of a prototype digital wallet that leverages blockchain's decentralized ledger, cryptographic principles, and smart contract functionalities.

Technological Aspects: The research delves into the technological intricacies of blockchain integration. Cryptographic protocols [1] play a pivotal role in securing transactions within digital wallets. This involves exploring encryption techniques, digital signatures, and hashing algorithms to safeguard user data and transactional integrity. The development of smart contracts further automates and secures payment processes, enhancing the efficiency of financial interactions within the metaverse.

Consensus Mechanisms: The research considers various consensus mechanisms employed in blockchain networks. By understanding and implementing suitable consensus mechanisms, such as proof-of-stake or proof-of-work, the research aims to ensure the integrity of the decentralized network, validate transactions, and prevent fraudulent activities within the metaverse.

Usability Studies: An integral aspect of the work involves conducting usability studies to assess the practicality and user-friendliness of the blockchain-integrated metaverse digital wallet. This phase aims to identify user preferences, address potential user experience challenges, and refine the design for optimal accessibility and acceptance.

The work description also encompasses the development of use cases for smart contracts within digital wallets, exploring scenarios such as automated payment processes, conditional transactions, and programmable agreements. These use cases aim to showcase the transformative potential of integrating blockchain technology into digital wallets, offering tangible benefits in terms of security, efficiency, and automation.

Throughout the work, a meticulous approach to security considerations is maintained. The integration of biometric authentication, decentralized identity management, and encryption mechanisms ensures a robust security framework that aligns with the demands of secure financial transactions within the metaverse.

The work description outlines a comprehensive plan for integrating blockchain technology into digital wallets, emphasizing the development of a prototype, exploration of technological aspects, implementation of consensus mechanisms, and usability studies. This multifaceted approach aims to contribute to the nascent but rapidly evolving field of secure digital transactions within the metaverse.

V. PROJECT EXECUTION PLAN

A. Project Tasks:

1. Literature Review: Conduct an extensive and methodologically rigorous literature review to comprehensively understand the current state of digital wallets, blockchain technology, and the metaverse. The review will adhere to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) methodology to ensure transparency, reproducibility, and academic rigor. This process includes formulating specific research questions, defining inclusion and exclusion criteria, performing structured searches across multiple databases, and synthesizing findings using a standardized screening and eligibility protocol.

By applying the PRISMA approach [24], the research will systematically identify existing gaps, challenges, and potential technological solutions in the domain of metaverse payment systems. The results of the literature review will serve as a critical foundation for the conceptual framework and prototype development phases. A PRISMA flow diagram will be included to illustrate the review process.

2. Conceptual Framework Development: Develop a robust conceptual framework based on the theoretical foundation established in the literature review. Define key terms and concepts related to secure digital wallets, blockchain, and the metaverse, aligning the framework with the objectives of the research.

3. Blockchain Integration Strategy: Define a strategic plan for integrating blockchain technology into digital wallets. Explore different blockchain architectures, consensus mechanisms, and cryptographic protocols [1] to inform the development of a secure and efficient digital wallet prototype that interoperates with metaverse platforms.

4. Prototype Development: Initiate the development of a prototype digital wallet, incorporating the proposed blockchain integration strategy with metaverse. This task involves coding the necessary functionalities, ensuring interoperability with metaverse platforms, and integrating security features such as biometric authentication and decentralized identity management.

5. Smart Contract Implementation: Implement and test smart contracts within the digital wallet prototype. Develop use cases for self-executing agreements, conditional transactions, and programmable functionalities, aiming to showcase the transformative potential of blockchain-based automation in metaverse payments.

6. Usability Studies: Conduct usability studies to evaluate the user experience of the developed digital wallet prototype. Gather feedback on user interactions, identify potential challenges, and iteratively refine the user interface and functionalities to enhance accessibility and acceptance.

7. Security Assessments: Perform rigorous security assessments of the blockchain-based digital wallet prototype integrated with metaverse platforms. This includes vulnerability testing, penetration testing, and analysis of encryption mechanisms to ensure the robustness of the security framework. Address any identified vulnerabilities and reinforce security measures.

8. Regulatory Analysis: Undertake a comprehensive analysis of existing regulatory frameworks relevant to metaverse transactions and blockchain-based digital wallets. Identify legal considerations, compliance requirements, and potential challenges that may impact the deployment and adoption of the proposed solution.

9. Finalization of Prototype: Based on the feedback from usability studies and security assessments, finalize the blockchain-based digital wallet prototype integrated with metaverse platforms. Ensure that the prototype aligns with the defined conceptual framework, meets security standards, and adheres to regulatory requirements.

10. Research Paper Development: Draft research papers detailing the theoretical foundation, conceptual

framework, methodology, findings, and contributions of the research. Prepare manuscripts for publication in reputable academic journals and conferences.

11. Contribution to Field: Summarize the research findings and contributions, emphasizing how the integration of blockchain into digital wallets addresses identified gaps and challenges in metaverse payment systems. Showcase the potential impact of the research on the field of secure digital transactions within the metaverse.

B. Timeline:

TABLE I. PROJECT TIMELINE

Period	Milestones
Sept 2024 – Jun 2025	Literature Review
Jul 2025 – Dec 2025	-Conceptual Framework Development -Research Paper Development and Paper submission
Jan 2026 – Mar 2026	Blockchain Integration Strategy
Avr 2026 – Dec 2026	Prototype Development
Jul 2026 – Dec 2026	Research Paper Development and Paper submission
Jan 2027 – Mar 2027	Smart Contract Implementation
Avr 2027 – Jul 2027	Usability Studies
Aug 2027 – Dec 2027	-Security Assessments -Research Paper Development and Paper submission
Jan 2028 – Mar 2028	Regulatory Analysis
Avr 2028 – Jun 2028	Finalization of Prototype
Jul 2028 – Dec 2028	Research Paper Development and finalisation of thesis

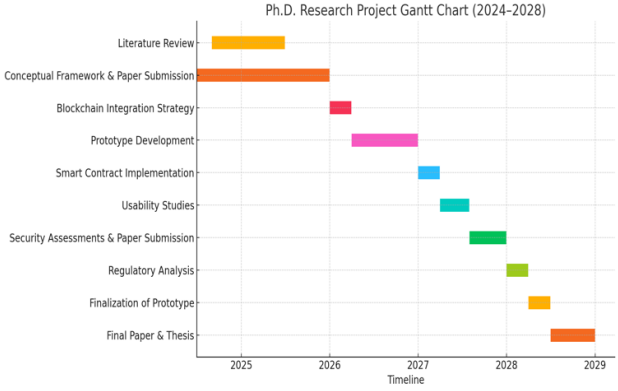


Fig. 1. GANTT Diagram.

C. Project Outputs:

The expected outputs of this research are multifaceted and align with the outlined tasks and objectives:

1. **Secure Digital Wallet Prototype**: The primary output is the development of a secure digital wallet prototype, showcasing the integration of blockchain technology into metaverse transactions. The prototype will demonstrate advanced security measures, smart contract functionalities, and usability improvements.

2. **Contributions to the Field**: The research aims to make substantial contributions to the field by addressing identified gaps and challenges. The expected impact includes advancing the understanding of secure digital transactions within the metaverse, proposing innovative solutions, and providing insights into the integration of blockchain technology.

3. **Potential for industry adoption and entrepreneurial opportunities**: The prototype and research findings not only have the potential for industry adoption but also create opportunities for entrepreneurship. The insights derived from this research can guide the development of commercial applications, influencing the design and deployment of secure digital wallets within the evolving metaverse landscape. This not only positions individuals to contribute to the industry but also opens pathways for entrepreneurial pursuits. The culmination of this project offers the researcher the prospect of transitioning from academic exploration to practical implementation, providing the groundwork to potentially establish a startup or embark on entrepreneurial ventures in the burgeoning domain of secure digital transactions within the metaverse.

In summary, the Project Execution Plan outlines a comprehensive strategy for conducting research, developing a prototype, and contributing to the academic and practical aspects of secure digital transactions within the metaverse. The timeline provides a realistic schedule, and the expected outputs align with the defined objectives, aiming for both academic excellence and practical impact.

VI. CONCLUSION AND PERSPECTIVES

This paper explored the integration of blockchain-based digital wallets within the metaverse, emphasizing the need for secure, transparent, and interoperable payment infrastructures. Through a comprehensive review of current literature and the design of a multi-phase research plan, the study highlighted critical challenges in existing systems and proposed a conceptual framework for a decentralized, user-centric financial solution.

The main contributions include a strategy for incorporating smart contracts, decentralized identity, and cryptographic protocols into a functional wallet prototype tailored to metaverse ecosystems. The proposed solution intends to not only address technical gaps but also considers usability, regulatory compliance, and long-term scalability.

Future work will focus on the specification, design, development and validation of the prototype within real-world virtual platforms, deeper analysis of legal frameworks, and the incorporation of privacy-enhancing technologies such as zero-knowledge proofs. These directions aim to strengthen trust and resilience in emerging digital economies.

ACKNOWLEDGMENTS

This work was supported by the Fundação para a Ciência e Tecnologia **(FCT)** within projects: **UIDB/04466/2020** and **UIDP/04466/2020**, and also in part by the project Blockchain.PT – Agenda Decentralize Portugal with Blockchain, (Project No 51), WP 7:Interoperability, call No 02/C05-i01.01/2022, funded by the Portuguese Recovery and Resilience Program (PPR), the Portuguese Republic and the European Union (EU) under the framework of Next Generation EU Program.

REFERENCES

[1] Zhou, Y., & Kapoor, K. (2023). Evolution of E-Wallets in Digital Finance. *Journal of Financial Innovation*, 8(1), 34–51.
[2] J. Lee et al. (2023). Security Architecture for E-wallets in Web3 Ecosystems. *IEEE Transactions on Dependable and Secure Computing*.
[3] ISO/TC 307. (2023). Blockchain and Distributed Ledger Technologies – Security Guidelines.
[4] Schaub, A. et al. (2022). Privacy-preserving payments in e-wallets: A usability analysis. *Privacy & Usability Journal*, 14(2), 88–102.
[5] Ali, M. et al. (2023). Metaverse Communications and Security. *IEEE Communications Surveys & Tutorials*.
[6] Lin, Y., & Chang, H. (2023). Digital Wallets in XR-based Commerce. *ACM SIGGRAPH Asia*.
[7] D. Brown & L. Green (2023). Trust and UX in Decentralized Wallets. *HCI for Financial Systems Conference*.
[8] Nakamoto, S. (2008). Bitcoin: A Peer-to-Peer Electronic Cash System.
[9] Christodoulou, K. et al. (2022). NFTs and the Metaverse Revolution. In *Blockchains and the Token Economy*.
[10] Wood, G. (2023). The Future of Smart Contracts in Web3. *Polkadot Whitepapers*.
[11] Ethereum Foundation. (2023). ERC-4337: Account Abstraction for Smart Wallets.
[12] W3C (2023). Decentralized Identifiers (DID) v1.0 Specification. https://www.w3.org/TR/did-core
[13] Huang, H. et al. (2024). Economic Systems in the Metaverse. *ACM Computing Surveys*, 56(4), Article 99.
[14] Belk, R. et al. (2022). Money and Ownership in the Metaverse. *Journal of Business Research*, 153, 198–205.
[15] Wang, H. et al. (2023). A Survey on the Metaverse. *IEEE IoT Journal*, 10(16), 14671–14688.
[16] Gatteschi, V., Lamberti, F., et al. (2022). The Rise of Mixed Reality Payments. *Elsevier Future Generation Computer Systems*, 138, 372–387.
[17] Choi, Y. et al. (2023). Blockchain Wallets for Virtual Economies. *IEEE Blockchain Transactions*.
[18] Johnson, M., & Smith, A. (2019). Decentralized Finance. *Journal of Financial Technology*, 5(2), 123–145.
[19] Financial Stability Oversight Council. (2022). *Regulatory Frameworks for Virtual Currencies*.
[20] Blockchain Institute. (2023). *Blockchain Integration for Immersive Economies*.
[21] Web3Auth (2023). Unified Key Management for Web3. https://web3auth.io
[22] Buterin, V. (2022). Soulbound Tokens: Decentralized Identity and Reputation. https://vitalik.eth.limo
[23] ZKProof Consortium (2024). Zero-Knowledge Proofs in Digital Identity. *ZK Privacy Standards Initiative*.
[24] Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... & Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Systematic Reviews*, 10(1), 89. https://doi.org/10.1186/s13643-021-01671-z
[25] Hassaan, M. (2025), "Understanding customer's intention to adopt metaverse banking services in Pakistan: a qualitative study", *Qualitative Research in Financial Markets*, Vol. 17 No. 3, pp. 661-683. https://doi.org/10.1108/QRFM-02-2024-0052

A Decentralised and Scalable Approach for Intrusion Detection in Cybersecurity Networks

José M. Franco-Valiente*, Jesús Calle-Cancho†, Carlos Cañada‡, Juan Mario Haut‡
*CETA-CIEMAT, Trujillo, Spain
†Dept. of Computing and Telematics Engineering, Univ. Extremadura, Cáceres, Spain
‡Dept. of Tech. of Computers and Communications, Univ. Extremadura, Cáceres, Spain
Emails: josemiguel.franco@ciemat.es, jesuscalle@unex.es, ccanadar@unex.es, juanmariohaut@unex.es

***Abstract*—The increasing volume and complexity of network traffic pose significant challenges for intrusion detection in cybersecurity systems. Traditional solutions often struggle to scale effectively in big data environments, prompting the need for distributed approaches. This paper presents a decentralised and horizontally scalable intrusion detection system built on Apache Spark. Unlike previous distributed IDS implementations, our approach explicitly addresses class imbalance and scalability on large datasets by integrating a five-stage processing pipeline and distributed XGBoost training. The system is evaluated using both the initial and an enhanced replicated version of the NSL-KDD dataset, reaching a memory footprint of 100.02 GB. Results show near-linear speedup as additional worker nodes are introduced, substantially improving scalability. Furthermore, we examine how class balancing with ADASYN mitigates detection gaps for minority classes (e.g., R2L, U2R). These findings underscore the feasibility of combining distributed computing and ensemble learning to tackle modern cybersecurity challenges in cloud-native environments.**

***Index Terms*—IDS, Distributed Machine Learning, Cybersecurity, Big Data, Cloud Computing**

I. Introduction

The exponential growth of Internet-based services and connected devices has fundamentally reshaped modern communication networks. Digital services span diverse sectors—including finance, healthcare, government, and industrial control systems—leading to unprecedented increases in network traffic volume, velocity, and heterogeneity [1].

Traditional Intrusion Detection Systems (IDS) have primarily relied on centralised, rule-based approaches, often built around fixed signature databases or heuristic rules. While suitable for static environments with modest data throughput, these systems struggle significantly with high-volume, high-velocity traffic, resulting in scalability limitations, high false-positive rates, and inadequate adaptability to emerging threats, including zero-day attacks. Zero-day attacks are cyber-attacks exploiting previously unknown vulnerabilities in software or hardware, for which no patch or defence currently exists [2]. These limitations strongly motivate the need for scalable, adaptive, and distributed solutions capable of effectively managing large-scale traffic and rapidly evolving threats, which is precisely what we propose here.

Given these challenges, our approach leverages distributed computing and advanced machine learning techniques to enhance intrusion detection capabilities.

II. Related Work

Recent research on distributed IDS has explored big data frameworks such as Hadoop, Spark, and Flink. Hadoop-based IDS improved batch-processing capabilities but lacked the efficiency necessary for near-real-time detection [3]. Spark addresses some of these shortcomings by supporting rapid iterative computations and native distributed memory management [4]. Nevertheless, many existing IDS solutions continue to encounter limitations arising from class imbalance, computational bottlenecks, or inadequate validation on large datasets.

Our work extends previous approaches by integrating Spark's distributed processing efficiency with XGBoost's ensemble learning capabilities. We explicitly incorporate adaptive class balancing and validate our system using datasets significantly larger than the original **NSL-KDD** [5]. Unlike the systems proposed in [6], we systematically assess scalability, communication overhead, and sensitivity to class imbalance.

III. Methodology

The proposed intrusion detection system utilises a structured five-stage pipeline implemented in **Apache Spark**, designed for deployment in distributed cloud environments. The first four stages execute sequentially, leveraging multi-threaded parallelism through libraries such as **numpy**, **pandas**, **scikit-learn**, and **imbalance-learn**. Only the final stage (model training) is executed entirely in a distributed manner on **Spark**.

A. Data Intake and Initial Processing

Raw network traffic data, originally in CSV format, is converted into **Apache Parquet** for its efficient columnar storage, enabling swift distributed querying and processing [7].

B. Dataset Description

We utilise the **NSL-KDD** [5] dataset, a refined adaptation of the classic KDD'99, which contains 125,973 records spanning 41 features (both continuous and categorical). A pronounced class disproportion exists, with the majority of records classified as *Normal* or *DoS*, while classes such as *R2L* and *U2R* are severely under-represented.

TABLE I
OVERVIEW OF THE ORIGINAL NSL-KDD DATASET

Attribute	Value
Total Records	125,973
Number of Features	41 (continuous and categorical)
Majority Classes	Normal, DoS
Minority Classes	R2L, U2R
Format	CSV (converted to Parquet)
Initial Size on Disk	~19 MB

After applying **ADASYN** [8], the dataset is balanced to roughly 67,000 instances per class (336,665 records in total). To evaluate the system's ability to manage an expanding volume of data—as encountered in many real-world operational scenarios—the dataset is subsequently replicated until it reaches 100.02 GB in memory, whilst preserving the original class proportions.

C. Statistical Analysis and Data Exploration

This step computes statistical metrics (means, medians, standard deviations), detects outliers via Z-score analysis, and produces correlation matrices to identify inter-feature dependence. These insights refine the data transformations and confirm the need for a rebalancing strategy.

TABLE II
CLASS DISTRIBUTION OF THE ORIGINAL NSL-KDD DATASET

	Train		Test	
Class	**# Records**	**% of Total**	**# Records**	**% of Total**
Normal	67,343	53.58%	9,711	43.07%
DoS	45,927	36.54%	7,460	33.10%
R2L	11,656	9.27%	2,885	12.80%
Probe	995	0.79%	2,421	10.74%
U2R	52	0.04%	67	0.30%
Total	**125,973**	**100.00%**	**22,544**	**100.00%**

D. Data Curation and Feature Engineering

Preprocessing operations included **min-max normalisation**, a technique that scales numerical features to a defined range, typically between 0 and 1, improving the performance of machine learning algorithms sensitive to feature scales. Additionally, we applied **one-hot encoding** to transform categorical variables into binary vectors, ensuring that categorical data could be effectively processed by our classification model. Integrity checks were also performed to detect and address missing values, inconsistent data types, or malformed records.

E. Class Imbalance Management

To address class imbalance, we applied **ADASYN** [8], which generates synthetic samples focused on minority classes, particularly those with complex boundaries. Specifically, we configured the ADASYN algorithm with a fixed `random_state=42` to ensure reproducibility across experiments. This method effectively balanced the dataset, achieving a near-equal distribution across all attack categories.

TABLE III
DISTRIBUTION OF RECORDS BY ATTACK CLASS (BALANCED DATASET)

Class	# Records	% of Total
Normal	67,343	20.00%
DoS	67,361	20.01%
U2R	67,336	20.00%
R2L	67,329	20.00%
Probe	67,296	19.99%
Total	**336,665**	**100.00%**

F. Dataset Replication

Following class balancing, we implemented a chunk-based replication strategy to scale up the dataset, simulating conditions typical in real-world big data scenarios. To maintain data variability and avoid exact duplication, we added small random noise ($\mu = 0$, $\sigma = 0.001$) to numeric features within each replicated chunk. This approach allowed us to scale the dataset to exceed 100 GB in memory, facilitating robust performance testing. Crucially, we ensured that the original balanced class distributions were preserved throughout the replication process.

TABLE IV
FINAL CLASS DISTRIBUTION (100+ GB DATASET)

Class	# Records	% of Total
DoS	20,881,290	20.01%
Normal	20,876,330	20.00%
Probe	20,875,400	20.00%
U2R	20,874,160	20.00%
R2L	20,862,690	19.99%
Total	**104,369,870**	**100.00%**

G. Distributed Model Training

The final stage involves distributed training using a **XGBoost** classifier integrated with **Apache Spark**. XGBoost is chosen for its resistance to overfitting, computational efficiency, and high performance on structured tabular data [9]. Alternatives like Random Forest and LightGBM were evaluated, but XGBoost delivered superior results in preliminary experiments [10].

IV. EXPERIMENTAL SETUP

A. Infrastructure and Deployment

Experiments were conducted on the CETA-CIEMAT cloud infrastructure, featuring eight dual-processor AMD EPYC 9845 servers. Spark clusters were deployed via Infrastructure as Code (IaC) using OpenTofu [11] ensuring full reproducibility.

- **Dataset:** NSL-KDD balanced dataset (~11 GB after replication)
- **Model:** XGBoost (Gradient Boosted Trees)
- **Cluster Configurations:** 1, 2, 4, 8, and 16 Spark workers
- **Metrics:** Accuracy, sensitivity, specificity (overall and per class), training time

B. Spark Configuration

TABLE V
INITIAL SPARK CONFIGURATION

Setting	Value
Spark Version	3.3.0
Deployment Mode	Standalone
Executor Memory	4 GB
Cores per Executor	4

C. Evaluation Criteria

We primarily focus on two dimensions:

- **Classification Performance**: Measured through accuracy, precision, and recall, providing insight into both overall detection capability and minority class detection.
- **Horizontal Scalability**: Assessed by observing training and inference times across varying cluster sizes. We quantify improvements via speedup factors and percentage reductions in training time compared to the baseline (single-node or sequential approach).

V. PERFORMANCE EVALUATION

In this section, we present a detailed analysis of the system's performance under different configurations and worker counts. We first examine the initial Spark setup, followed by the optimised configuration, highlighting speedup factors and classification metrics. We then discuss the overall classification performance range and specific class-wise results.

A. Training Time and Speedup

Table VI shows the training time (in seconds) and speedup. Increasing workers notably reduces training time, achieving up to a 7.42x speedup.

TABLE VI
TRAINING TIME AND SPEEDUP ANALYSIS

Workers	Time (s)	Speedup	Accuracy	Avg Sens	Avg Spec
1	1768.23	1.00x	0.7642	0.6106	0.9263
2	1007.27	1.76x	0.7630	0.6040	0.9259
4	572.05	3.09x	0.7566	0.5927	0.9239
8	392.96	4.50x	0.7613	0.6061	0.9254
16	238.19	7.42x	0.7563	0.5984	0.9240

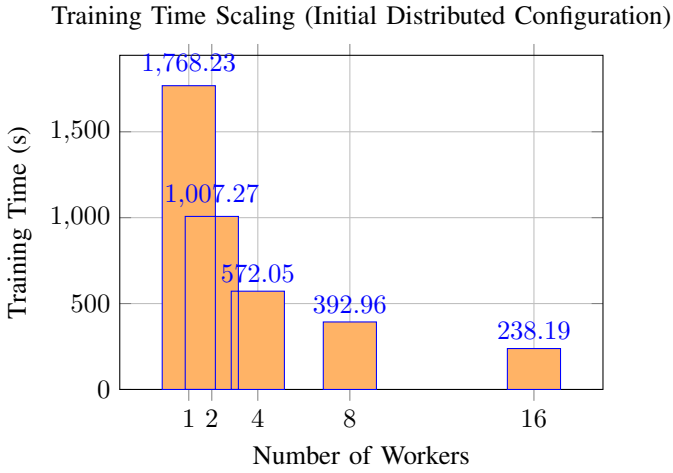


Fig. 1. Training time vs. Number of Spark workers (Initial Distributed Configuration).

B. Prediction Time Scalability

Besides training time, another crucial factor in real-world deployments is the prediction or inference time. We evaluate how prediction time scales with increasing Spark worker counts. Table VII summarises the prediction times (in seconds) and corresponding speedups for the distributed version.

TABLE VII
PREDICTION TIME SCALING ANALYSIS

Workers	Prediction Time (s)	Speedup
1	0.09	1.00x
2	0.10	0.90x
4	0.09	1.00x
8	0.09	1.00x
16	0.10	0.90x

Prediction time remains consistently low across all configurations, between 0.09 and 0.10 seconds. This confirms that **real-time inference is computationally inexpensive** and does not require large clusters.

C. Overall Classification Metrics

TABLE VIII
OVERALL CLASSIFIER PERFORMANCE

Version	Accuracy	Avg Sens	Avg Spec
Sequential	0.7642	0.6106	0.9263
Distributed	0.7563–0.7613	0.5927–0.6061	0.9239–0.9254

D. Class-Specific Performance

Finally, Table IX shows the sensitivity and specificity for each class when using a single worker as the baseline. We observe strong performance for Normal and DoS, while R2L and U2R remain challenging. The application of ADASYN helps mitigate these detection gaps in multi-worker setups, albeit not entirely eliminating them.

TABLE IX
CLASS-SPECIFIC PERFORMANCE (BASELINE: 1 WORKER)

Attack Type	Sensitivity	Specificity
DoS	0.7066	0.9857
Normal	0.9649	0.7220
Probe	0.8401	0.9386
U2R	0.3582	0.9987
R2L	0.1830	0.9863

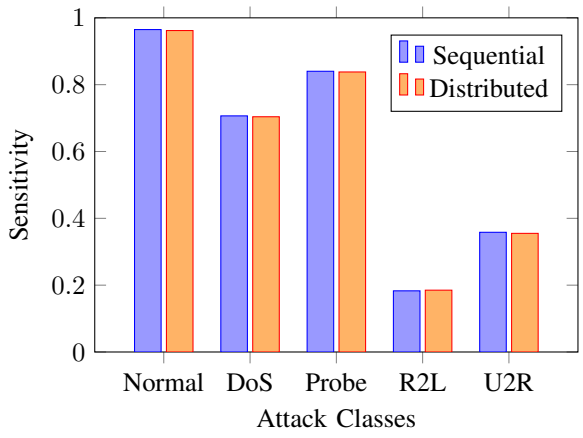


Fig. 2. Class wise Sensitivity: Sequential vs. Distributed vs. Dist. Optimised.

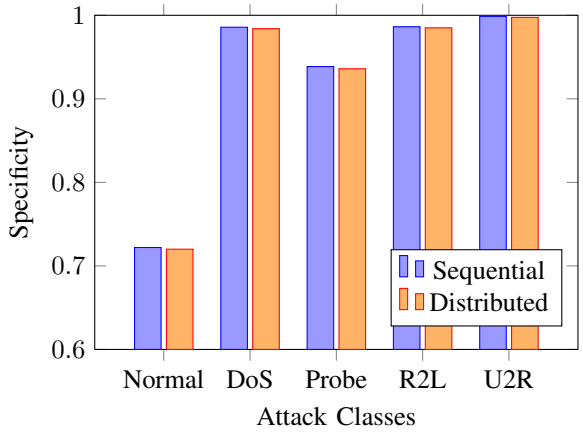


Fig. 3. Class wise Specificity: Sequential vs. Distributed

E. Discussion of Results

Overall, these findings highlight near-linear gains in training speedup and the robustness of classification metrics across the initial Spark distributed configuration. Despite the overhead from shuffle operations and serialisation, the system achieves significant reductions in training time. Furthermore, while minority classes (U2R, R2L) remain challenging, ADASYN oversampling helps narrow detection gaps.

VI. PRACTICAL IMPLICATIONS

These findings confirm that the proposed IDS can effectively operate in large-scale, real-time contexts. The system's scalability ensures adaptability to varying workloads, while minority-class sensitivity—though not perfect—is noticeably improved. Such an approach is therefore well-suited to domains like critical infrastructure or financial services, where detection of rare yet severe attacks is paramount.

VII. CONCLUSIONS

The proposed scalable IDS demonstrates robust horizontal scalability, high classification accuracy, and adaptability to class imbalance. Key observations include:

- A near-linear reduction in training time at moderate node counts.
- Improved detection rates for under-represented attack classes following ADASYN.
- Overheads from shuffle and communication remain manageable.

A. Limitations and Future Work

While results are promising, replicating a large dataset and applying ADASYN can be resource-intensive. Moreover, this study chiefly focuses on NSL-KDD, leaving newer threat vectors and encrypted traffic for subsequent exploration. Future research directions include:

1) **Exploring Alternative Balancing Techniques**: Such as SMOTE [12] variations or cost-sensitive methods.
2) **Optimising Spark Configurations**: Tuning shuffle parameters and leveraging advanced scheduling to reduce communication overhead.
3) **Adopting Online Learning**: To handle constantly evolving threats in real-time streaming scenarios.

ACKNOWLEDGEMENTS

This work is funded by Project C109/23 "Strategic Project UEx (Polytechnic School of Cáceres) - INCIBE". Also, this work was enabled by the computing facilities of the Extremadura Research Centre for Advanced Technologies (CETA-CIEMAT), funded by the European Regional Development Fund (ERDF). The authors gratefully acknowledge the support and resources provided.

REFERENCES

[1] Z. Ahmad, A. S. Khan, C. W. Shiang, J. Abdullah, and F. Ahmad, "Network intrusion detection system: A systematic study of machine learning and deep learning approaches," *Expert Systems*, vol. 32, Oct 2020. [Online]. Available: https://doi.org/10.1002/ett.4150

[2] H.-J. Liao, C.-H. R. Lin, Y.-C. Lin, and K.-Y. Tung, "Intrusion detection system: A comprehensive review," *Journal of Network and Computer Applications*, vol. 36, pp. 16–24, Jan 2013. [Online]. Available: https://doi.org/10.1016/j.jnca.2012.09.004

[3] S. S. Dhaliwal, A.-A. Nahid, and H. Abbas, "Effective intrusion detection system using xgboost," *Information*, vol. 9, no. 7, p. 149, Jul 2018. [Online]. Available: https://doi.org/10.3390/info9070149

[4] K. Yogesh, M. Karthik, T. Naveen, and S. Saravanan, "Design and evaluation of scalable intrusion detection system using machine learning and apache spark," in *2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*, 2019, pp. 1–7.

[5] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the kdd cup 99 data set," in *2009 IEEE symposium on computational intelligence for security and defense applications*. Ieee, 2009, pp. 1–6.

[6] M. Elbakri, S. Abdellatif, M. Sayed, and M. Ahmed, "Real-time network intrusion detection systems using apache spark," *Cluster Computing*, vol. 21, pp. 185–199, 2018. [Online]. Available: https://doi.org/10.1007/s10586-017-0956-5

[7] A. S. Foundation, "Apache parquet: Columnar storage," https://parquet.apache.org/, accessed: 2025-03-27.

[8] H. He, Y. Bai, E. A. Garcia, and S. Li, "Adasyn: Adaptive synthetic sampling approach for imbalanced learning," *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN)*, vol. 31, pp. 1322–1328, Jun 2008. [Online]. Available: https://doi.org/10.1109/IJCNN.2008.4633969

[9] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, 2016, pp. 785–794. [Online]. Available: https://doi.org/10.1145/2939672.2939785

[10] G. P. Gupta and M. Kulariya, "A framework for fast and efficient cyber security network intrusion detection using apache spark," *Elsevier BV*, vol. 93, pp. 824–831, 01 2016. [Online]. Available: https://doi.org/10.1016/j.procs.2016.07.238

[11] OpenTofu Community, "OpenTofu: An Open Source Terraform Alternative," Available at https://opentofu.org/, accessed: 2025-03-27.

[12] D. Elreedy and A. F. Atiya, "A comprehensive analysis of synthetic minority oversampling technique (smote) for handling class imbalance," *Information Sciences*, vol. 505, pp. 32–64, Jul 2019. [Online]. Available: https://doi.org/10.1016/j.ins.2019.07.070

CyberChatbot: A RAG-based Chatbot to Simplify Cybersecurity Regulatory Compliance in Spanish

1st Víctor González
2nd Belén María Ramírez
Media Engineering Group
University of Extremadura
Cáceres, Spain
{victorgomo, belramirez}
@unex.es

3rd Alberto Fernández-Isabel
Data Science Laboratory
Rey Juan Carlos University
Móstoles (Madrid), Spain
alberto.fernandez.isabel@urjc.es

4th Mohammadhossein Homaei
Media Engineering Group
University of Extremadura
Cáceres, Spain
mhomaein@alumnos.unex.es

5th Mar Ávila
6th Andrés Caro
Media Engineering Group
University of Extremadura
Cáceres, Spain
{mmavila, andresc}@unex.es

***Abstract*—Understanding cybersecurity regulatory compliance can be complex, particularly for non-expert users and organizations operating in Spanish-speaking contexts. This paper presents *CyberChatbot*, an intelligent chatbot system built using Retrieval-Augmented Generation (RAG) architecture to simplify and explain cybersecurity regulations in Spanish. The system leverages a curated knowledge base of legal and regulatory texts. It combines this information with state-of-the-art natural language processing techniques to provide accurate, context-aware, and user-friendly query responses. Designed to bridge the gap between technical legal language and everyday comprehension, *CyberChatbot* enhances accessibility to the legal cybersecurity domain by translating dense regulatory information into clear, actionable insights. The chatbot supports dynamic interactions, enabling users to ask follow-up questions and receive tailored explanations. Evaluation results demonstrate the system's effectiveness in improving users' understanding and confidence when navigating cybersecurity requirements. Therefore, the proposal represents a step forward in democratizing access to digital policy through Natural Language Processing (NLP) technologies.**

***Index Terms*—Large Language Models, Cybersecurity, Regulatory Compliance, Retrieval-Augmented Generation, Information Retrieval**

I. Introduction

The rapid growth of digital services, the widespread adoption of interconnected devices, and the escalating frequency and complexity of cyber threats have elevated cybersecurity to a top priority for governments, businesses, and society as a whole [1]. As digital ecosystems become increasingly complex, the risks associated with data breaches, system vulnerabilities, and malicious attacks continue to intensify, demanding more proactive and comprehensive security measures.

In response to this evolving threat landscape, numerous regulatory agencies have emerged (e.g., National Institute of Standards and Technology (NIST), International Electrotechnical Commission (IEC), and International Organization for Standardization (ISO)) at both national and international levels to enforce rigorous cybersecurity standards across a variety of sectors. Among the most prominent are international norms such as ISO/IEC 27001 [2], which outlines best practices for information security management; European regulations, such as the NIS 2 Directive to ensure a high common level of cybersecurity in the European Union [3]; national regulations like the Spanish National Security Scheme (ENS) [4], which establishes mandatory cybersecurity criteria for public sector entities in Spain; and globally influential guidelines such as NIST SP 800-53 [5], which provides a catalog of security and privacy controls for federal information systems. While these frameworks are fundamental to strengthening cybersecurity posture, their implementation presents significant challenges. These challenges stem primarily from the technical complexity, extensive documentation, and often, legalistic language that characterize such regulations.

This barrier is especially strong for Small and Medium-sized Enterprises (SMEs), local governments, educational institutions, and other organizations with limited resources. Many of them do not have cybersecurity experts or legal staff, so they face difficulties when trying to understand and apply technical regulations. For this reason, there is a strong need for practical tools that can help translate cybersecurity laws into clear and useful information for everyday use.

The main motivations for this work are:

- Regulatory complexity: Cybersecurity regulations such as ISO/IEC 27001, ENS, and NIS 2 Directive use technical and legal language that is hard to understand for many users.
- Limited resources in organizations: Many small institutions lack cybersecurity or legal experts and need tools that can support them without requiring advanced knowledge.
- Language gap: Most available Artificial Intelligence (AI) tools focus on English and do not provide full support for Spanish-speaking users or local Spanish regulations.
- Poor accessibility of official documents: Regulations are often long PDF files, which are hard to search or interpret quickly.
- Recent progress in AI: Advances in large language models and Retrieval-Augmented Generation (RAG) techniques make it possible to create systems that give

accurate and document-based answers.
- High demand for guidance: Many professionals and institutions need better support to follow cybersecurity rules correctly and confidently.

To respond to these challenges, this paper presents *CyberChatbot*, an intelligent chatbot system built to support Spanish-speaking users in understanding and applying cybersecurity regulations. The chatbot answers questions in natural language, using official regulatory documents as its main source of information.

The core of the system lies in a combination of advanced Large Language Models (LLMs) for natural language understanding and generation, integrated with RAG [6] techniques implemented via the Langchain framework [7]. This architecture enables the system to dynamically retrieve relevant content from a curated corpus of official regulatory documents and generate precise, trustworthy responses grounded in those sources.

A set of experiments has been conducted in which different LLMs were tested within the same pipeline to validate the effectiveness of this approach. Retrieved passages were supplied to each model to generate answers, which were evaluated across several critical dimensions, including accuracy, traceability to source documents, and clarity of the generated responses. These evaluations aim to determine the most suitable LLM for deployment in real-world, Spanish-language regulatory environments.

The remainder of this paper is structured as follows. Section II discusses related work in the regulatory chatbots domain and compliance-focused Natural Language Processing (NLP) systems. Section III describes the proposed architecture and development of the *CyberChatbot* system. Section V addresses the conducted experiments, while Section VI concludes and proposes future guidelines.

II. Related work

Recently, AI-driven chatbots have been developed to support people in various contexts, from mental health assistance to education and customer service. This trend has been significantly accelerated by the emergence of large language models (LLMs) such as GPT-4, PaLM, and Claude, demonstrating a high ability to understand and generate human-like language across various topics.

As a result, AI-driven chatbots are replacing more rigid, rule-based systems and being integrated into existing platforms, offering on-demand support at scale. This fact reshapes how users access information, receive assistance, and interact with complex systems through natural language [8].

In the cybersecurity domain, novel efforts have begun to explore the possibilities of these systems. These specialized chatbots assist users in understanding security threats, guiding them through best practices, and responding to incidents in real time. For instance, they can help users identify phishing attempts [9], manage password hygiene [10], or respond to suspicious activities [11] by providing automated, context-aware recommendations.

In particular, some of these chatbots have been developed to help people understand and follow cybersecurity standards. These tools are useful for companies and professionals who need to follow rules like ISO/IEC 27001, NIST SP 800-53, or the ENS.

Some of the most important AI-driven chatbots focus only on ISO standards. For example, Experta [12], created by Advisera, is trained with trusted content from ISO standards like ISO/IEC 27001, ISO 9001, and ISO 14001. It can answer questions about how to apply these standards.

Another example is ISO 27001 Copilot [13], made by Better ISMS. It is an AI-driven chatbot that helps people apply and maintain an Information Security Management System (ISMS) based on ISO 27001.

A different project is Botable Compliance Chatbot [14], which helps employees follow their company's internal policies and procedures. This chatbot can be adapted to include different standards, including ISO, if the documents are uploaded.

There is also an open-source chatbot called Security Docs Guide Chatbot [15], which anyone can use and modify. It can be trained with documents like ISO controls, NIST guides, or ENS laws.

To conclude, AI-driven chatbots already address the topics related to cybersecurity compliance, especially with ISO 27001. These systems are useful for understanding and applying standards correctly. However, there is no public chatbot that supports ISO, NIST, and ENS in the same system or gives full support in Spanish. Therefore, this work aims to provide a practical solution that combines these standards and supports Spanish-speaking users in resolving their questions about cybersecurity laws and regulations.

III. Proposal

In this work, we propose the development of a AI-driven chatbot system called *CyberChatbot* designed to support Spanish-speaking users with questions about cybersecurity regulatory compliance. The chatbot uses a RAG approach and is trained with official documents related to three important frameworks: ISO/IEC 27001, NIS 2 Directive (EU) 2022/2555, and the ENS from Spain.

The main goal of the AI-driven chatbot is to help users better understand the content of these standards. Users can ask informal questions such as "What does ISO 27001 say about access control?" or "Does my system meet ENS requirements?" and they can obtain a clear and helpful answer from the system, based on the official documents. The system is focused on making complex legal or technical texts easier to understand.

Unlike other chatbots that support only one framework or answer in English, our system combines multiple standards and offers full support in Spanish. This work aims to provide a real solution for professionals, students, and institutions to support applying cybersecurity regulations in their daily work.

The proposal, as presented in the methodology section, can be easily adapted to other regulatory compliance documents

from other countries and in other languages or even to other types of documents with other purposes.

IV. Methodology

This section details the complete workflow used to build and evaluate the proposed chatbot system. It covers the selection and preprocessing of legal documents, the implementation of the retrieval mechanism, the RAG-based system architecture, and the integration of multiple generative models for comparative analysis.

A. Regulatory dataset

The methodology starts by assembling a dataset composed of three foundational cybersecurity regulations that are widely recognized within Spain and the European Union to support the development and evaluation of the chatbot:

- Royal Decree 311/2022 (Spanish National Security Scheme) – This national regulation defines the security policy and minimum requirements for public sector IT systems in Spain, establishing risk-based management and control measures.
- UNE-EN ISO/IEC 27001:2017 – The international standard for Information Security Management Systems (ISMS), which includes guidelines on risk assessment, control implementation, and continuous improvement.
- Directive (EU) 2022/2555 (NIS 2 Directive) – A European directive that updates the legal framework for cybersecurity across the EU, extending its scope and enforcement mechanisms to ensure a high common level of cybersecurity across sectors.

This selection captures the diversity of cybersecurity regulation in scope and application, spanning legislative levels (national and European), document types (legal acts and technical standards), and target sectors (public administration and critical infrastructure operators).

B. Document retrieval and preparation

Each regulatory document was first preprocessed and transformed into a structured format suitable for downstream processing to enable the chatbot to provide accurate and grounded answers based on real legal texts. The documents in the regulatory corpus were originally provided in PDF format, which poses significant challenges for structure-aware processing due to the complexity and variability of legal layouts. These documents often include nested lists, footnotes, multi-column sections, tables, and detailed tables of contents, making naive text extraction insufficient.

A preprocessing pipeline has been developed to extract and export the content of each document into both Markdown and HTML formats while preserving the logical and hierarchical structure of the original text to address these issues. This process required multiple iterations and versions of the pipeline, progressively improving the quality of the extracted content through refinements in structure recognition and formatting. By doing so, the transformed documents must retain the semantic coherence, enabling effective alignment with the language model input and supporting accurate document retrieval within the chatbot system. A schematic overview of this pipeline is shown in Figure 1.

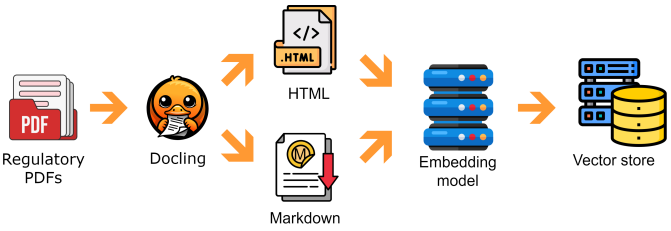


Fig. 1. Pipeline for preprocessing regulatory documents using Docling.

The core tool used in the process is *Docling* [16], a document ingestion and transformation framework designed to parse diverse types of documents to make them ready for Generative AI workflows. In this pipeline, each PDF is first parsed and loaded using the `DoclingLoader` class. This export process not only simplifies further processing but also improves document readability and traceability.

In addition to the *Docling* pipeline, subsequent versions of the system incorporated the use of `pdfplumber`, a layout-aware PDF parser, to improve the handling of complex visual elements, such as multi-level tables of contents, nested lists, and embedded tables. These structures are common in legal and normative texts but are often misrepresented or lost when using generic text extraction tools. Thus, working together, these tools achieved a significantly more accurate and interpretable output.

At the end of this preprocessing stage, the enhanced version allowed for more faithful preservation of tabular data and document hierarchy, which are essential for maintaining the legal context of the source material. The enriched outputs were integrated into the Markdown and HTML export routines, resulting in representations that retained both the form and function of the original documents. This preprocessing step was critical to ensure that the downstream AI components, including language models and retrievers, interacted with high-quality, structurally sound representations of the normative texts.

This document transformation stage was essential for ensuring that the chatbot system could work with text representations that closely reflect the original legal structure of the source documents.

C. System RAG Architecture

The structure of the RAG system is shown in Figure 2. When a user submits a question, the system first converts the query into embeddings using the model *bge-m3* [17]. These embeddings are then used to search a *FAISS* vector database [18], which contains the pre-processed content of official documents. The system retrieves the 3 most similar document segments based on similarity.

This context is combined with the original user question and inserted into a predefined prompt template, which guides

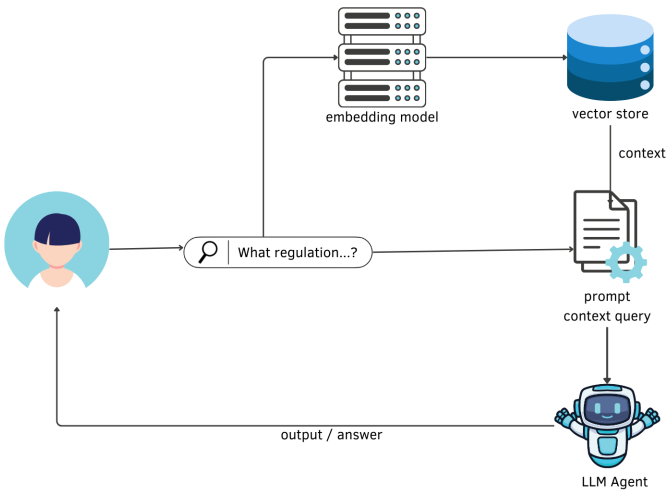


Fig. 2. RAG Architecture

the behavior of the language model. The prompt and retrieved context are sent to the language model interface, which allows communication with multiple LLMs. This design enables the comparison of different language models under the same conditions to evaluate their performance.

The system follows a RAG approach. It separates the search and generation steps to ensure that the answer is based only on the relevant parts of the documents. The modular structure also makes it easier to experiment with different components, such as changing the embedding model, the retrieval algorithm, or the LLM model.

Finally, the answer is returned to the user through a simple interface. This architecture helps to provide accurate, explainable, and standards-based answers, with the flexibility to adapt to different models and evaluation scenarios.

D. Generative models

Four LLMs, each with distinct architectural and training characteristics, have been selected to compare the proposal´s effectiveness within a cybersecurity legal context. The models are:

- *GPT-4o-mini* [19] – A lightweight variant of OpenAI's GPT-4o, optimized for fast response times and reduced computational cost while maintaining high performance in multilingual and legal reasoning tasks.
- *DeepSeek-R1:8B* [20] – A general-purpose open-source model trained on a diverse corpus, known for its robustness in knowledge-intensive tasks and efficient generation.
- *Gemma3:4B* [21] – A transformer-based model fine-tuned for instruction following and dialogue, designed for high contextual sensitivity and semantic coherence in domain-specific settings.
- *Mistral-Instruct-v0.3* [22] – A performant and compact open-weight model from the Mistral family, well-suited for retrieval-augmented pipelines due to its strong few-shot capabilities and competitive quality-to-size ratio.

All models were integrated into the same RAG pipeline and received identical retrieved document contexts to ensure a fair comparison. The objective was to identify the most suitable model for accurate, legally grounded, and user-accessible responses.

V. Experiments

This section details the experiments to evaluate the quality of the proposed *CyberChatbot* system. The evaluation was designed to measure the ability of different LLMs to generate accurate, grounded, and useful responses to cybersecurity-related legal questions based on official regulatory texts.

A. Question Design and Dataset

A set of 30 questions was manually crafted from three regulatory sources to build a representative benchmark. Each document contributed 10 questions, selected to reflect a range of diverse complexity, topics, and wording styles. The questions cover several aspects, such as obligations, definitions, implementation details, and compliance requirements, as summarized in Table I.

B. Model Evaluation Setup

As described in Section IV-D, four LLMs were selected for evaluation: *GPT-4o-mini*, *DeepSeek-R1*, *Gemma3*, and *Mistral v0.3*. These models were selected based on their architectural diversity, public availability, and relevance to retrieval-augmented generation tasks in constrained domains such as cybersecurity compliance.

Each question from the benchmark dataset was submitted to a shared Retrieval-Augmented Generation (RAG) pipeline, where the same retrieved context passages were provided to each model. All responses were generated using identical prompt structures and decoding parameters across models. This fact ensured a fair comparison.

For each model's output, three key metrics were evaluated:

- *Precision*: the proportion of correct answers among all model outputs.
- *Recall*: the proportion of relevant answers correctly retrieved and explained by the model.
- *F1 Score*: the harmonic mean of Precision and Recall, providing a balanced measure that accounts for both false positives and false negatives. It is useful when the class distribution is imbalanced or when both precision and recall are important.

C. Results and Discussion

The results revealed important differences in model performance depending on the source document and the underlying model architecture. This section discusses the key findings drawn from the metrics presented in Tables II and III, as well as Fig. 3, Fig. 4, and Fig. 5.

TABLE I
NUMBERED QUESTIONS GROUPED BY REGULATION

Question ID	Document	Question
ISO-1	ISO/IEC 27001	What requirements are imposed by section 6.1.2 for information security risk assessment?
ISO-2	ISO/IEC 27001	What elements must a 'Statement of Applicability' contain according to 6.1.3 d)?
ISO-3	ISO/IEC 27001	What documented information is required regarding information security objectives (6.2)?
ISO-4	ISO/IEC 27001	What aspects should be covered in the management review according to section 9.3?
ISO-5	ISO/IEC 27001	What guidelines are set for creating and updating documented information (7.5.2)?
ISO-6	ISO/IEC 27001	What is meant by operational control in the context of section 8.1?
ISO-7	ISO/IEC 27001	What competencies must the organization ensure according to section 7.2?
ISO-8	ISO/IEC 27001	How should the scope of the ISMS be established in accordance with section 4.3?
ISO-9	ISO/IEC 27001	What must the information security policy include according to clause 5.2?
ISO-10	ISO/IEC 27001	What corrective actions are required in the event of a nonconformity according to section 10.1?
ENS-1	ENS (BOE-A-2022-7191)	What does Article 17 establish regarding access control to information systems?
ENS-2	ENS (BOE-A-2022-7191)	How is the concept of 'minimum privilege' defined in Article 20 of the ENS?
ENS-3	ENS (BOE-A-2022-7191)	What procedures must be followed according to Article 33 in the event of a security incident?
ENS-4	ENS (BOE-A-2022-7191)	What criteria are used to determine the security category of a system according to Annex I?
ENS-5	ENS (BOE-A-2022-7191)	What does Article 28 mention about using compensatory measures instead of the established ones?
ENS-6	ENS (BOE-A-2022-7191)	How should the security audit be documented according to Annex III?
ENS-7	ENS (BOE-A-2022-7191)	What role does the Sectoral Committee for Electronic Administration play in Article 32?
ENS-8	ENS (BOE-A-2022-7191)	What document must include the adopted security measures according to Article 28?
ENS-9	ENS (BOE-A-2022-7191)	What examples of physical protection measures are included in Annex II?
ENS-10	ENS (BOE-A-2022-7191)	What does Article 3 say about systems that process personal data in relation to the GDPR?
NIS-1	Directive EU 2022/2555	What size criteria must companies meet to be considered within the scope of the Directive under Article 2?
NIS-2	Directive EU 2022/2555	What must national cybersecurity strategies include?
NIS-3	Directive EU 2022/2555	What functions are assigned to CSIRTs under Article 12?
NIS-4	Directive EU 2022/2555	What mechanisms does the Directive establish for real-time notification of significant incidents?
NIS-5	Directive EU 2022/2555	What measures must essential entities take against cyber threats according to Article 21?
NIS-6	Directive EU 2022/2555	How is cooperation between Member States coordinated through single points of contact (Article 10)?
NIS-7	Directive EU 2022/2555	What role does ENISA play according to Article 18 of the Directive?
NIS-8	Directive EU 2022/2555	What special provisions are established for data centers that are not part of the cloud?
NIS-9	Directive EU 2022/2555	What should CSIRTs do if they receive a vulnerability report from a researcher?
NIS-10	Directive EU 2022/2555	When is an incident considered significant?

TABLE II
PERFORMANCE COMPARISON BY DOCUMENT

Document	Precision	Recall	F1 Score
ISO27001	**0.8286**	0.8529	**0.8406**
ENS	0.6857	0.8276	0.7500
NIS 2	0.4211	**0.8889**	0.5715

1) Performance by document: Firstly, results aggregated by document are analyzed. Thus, how each regulatory source affected the overall performance of the models was considered. Table II summarizes the precision, recall, and F1 scores obtained when evaluating responses grouped by regulatory text. This analysis allows examining whether specific characteristics of each document—such as structure, linguistic complexity, or normative density—influence the models' ability to retrieve and generate accurate answers.

The ISO/IEC 27001 questions yielded the highest scores overall with an F1 score of 0.8406, followed by ENS, which achieved an F1 score of 0.7500. The lowest performance was observed for the NIS 2 Directive, where the F1 score dropped to 0.5715. This suggests that models found it easier to extract and summarize information from ISO27001, likely due to its more structured and procedural language. The high recall on NIS 2 (0.8889) combined with low precision (0.4211) indicates that models tended to retrieve large amounts of content but struggled to pinpoint specific answers, possibly due to the directive's more legalistic and abstract language.

A closer examination reveals that ISO/IEC 27001's modular architecture and consistent clause numbering created a strong alignment between question scope and textual structure, allowing models to perform both retrieval and answer extraction with high accuracy.

In contrast, ENS presented a mixed landscape: while its legal format provided well-defined sections, its hybrid nature—combining technical controls, legal terminology, and procedural content—introduced ambiguity that often resulted in overinclusive answers.

The NIS 2 directive posed the most significant challenge, not only because of its dense legislative prose but also due to the distribution of relevant information across multiple articles and recitals. Questions often require synthesizing scattered content or interpreting institutional roles and obligations that are not localized in a single passage. As a result, models frequently identified the correct general area of the directive but failed to isolate concise, definitive responses.

These findings highlight the impact of document structure and linguistic style on the effectiveness of automated question-answering systems and suggest that future improvements in retrieval-augmented models may require tailored preprocessing strategies for regulatory and legal texts.

2) Performance by model: The comparison of individual model performance across the evaluation dataset reveals distinct strengths and limitations in how each system approaches the answering task (see Table III). Among the four models assessed—*gpt4o-mini*, *DeepSeek-r1:8B, Gemma3:4B*, and *Mistral-Instruct-v0.3*—there are notable differences in their balance between precision, recall and F1 scores, which directly influence their F1 scores and, consequently, their overall

TABLE III
PERFORMANCE COMPARISON BY MODEL

Model	Precision	Recall	F1 Score
gpt4o-mini	**0.6957**	0.6957	0.6957
DeepSeek-r1:8B	0.6000	**1.0000**	0.7500
Gemma3:4B	0.5926	0.8421	0.6957
Mistral-Instruct-v0.3	0.6786	0.9048	**0.7755**

reliability in legal and technical domains.

GPT4o-mini displays symmetrical behavior across all three metrics, achieving a precision, recall, and F1 score of 0.6957. This balance suggests a stable, if moderate, performance where the model retrieves relevant content with a relatively low level of noise, but without reaching particularly high recall or specialization. Its consistent output implies a conservative retrieval strategy, avoiding overgeneralization but allevelso potentially missing more diffuse or inferential content.

DeepSeek-r1:8B stands out with a perfect recall of 1.0000, meaning it successfully retrieved relevant information for every question. However, this comes at the cost of precision, which drops to 0.6000, resulting in an F1 score of 0.7500. This extreme asymmetry suggests an aggressive retrieval approach, where the model tends to return large spans of text, ensuring that answers are included but often accompanied by excessive or tangential information. While this behavior may be advantageous in exploratory or open-ended contexts, it presents challenges for applications where concise and targeted responses are critical, such as compliance or policy validation.

Gemma shows a similar pattern to DeepSeek, though with slightly more balanced values. With a precision of 0.5926 and recall of 0.8421, its F1 score also reaches 0.6957, identical to that of GPT. This indicates that while Gemma is more inclined toward expansive recall than GPT, it does not achieve the same coverage as DeepSeek and still struggles with precision. The model seems to partially prioritize coverage of the semantic space but lacks sufficient filtering to produce focused, extractive responses.

Mistral emerges as the best-performing model in this comparison, achieving the highest F1 score of 0.7755. Its recall is strong at 0.9048, while maintaining a relatively high precision of 0.6786. This balance suggests that Mistral is capable of identifying relevant content across a wide range of contexts while also limiting the inclusion of irrelevant material. The model appears well-calibrated for regulatory question answering tasks, where the ability to capture distributed normative references must be complemented by a capacity to isolate precise information. Mistral's performance indicates a better understanding of document structure and a more effective internal filtering mechanism for determining answer boundaries.

Taken together, these results highlight that high recall alone is not a sufficient indicator of model effectiveness in legal and technical domains. Precision plays a decisive role in ensuring that retrieved content aligns closely with user expectations, especially in contexts where interpretive ambiguity is high. Models like GPT and Mistral, which achieve a strong balance between these metrics, appear better suited for deployment in environments where both completeness and accuracy are essential.

3) Performance by model and document: While analyzing model performance in aggregate provides useful insights, examining the interaction between each model and each regulatory document offers a more nuanced understanding of their behavior across different textual environments. The heatmaps in Fig. 3, Fig. 4, and Fig. 5 illustrate the precision, recall, and F1 scores obtained by each model across the three normative sources.

A striking pattern emerges in the recall matrix: DeepSeek consistently achieves a perfect recall score of 1.0000 across all documents. This confirms that it is highly effective at retrieving relevant content regardless of the regulatory context. However, this strength comes with a notable trade-off in precision, particularly in the NIS 2 Directive, where DeepSeek's precision drops sharply to 0.3000. Gemma exhibits a similar trend—maintaining perfect or near-perfect recall in NIS 2 and ENS, but with limited precision (also 0.3000 in NIS 2), suggesting a tendency to over-retrieve in documents with dispersed and legally dense content. In contrast, Mistral maintains a better balance across the board, demonstrating both high recall and relatively strong precision in all three cases. For ISO/IEC 27001 and ENS, Mistral reaches an F1 score of 0.8235, matching or exceeding the performance of other models, while also achieving the best balance for the NIS 2 Directive, where both recall and precision are elevated compared to the rest.

GPT's behavior is more conservative, particularly in ENS and NIS 2, where both recall and precision are moderate. Its precision is highest in ISO/IEC 27001, at 0.8750, which aligns with the document's structured nature and suggests that GPT benefits from modular regulatory texts with well-scoped clauses. However, GPT struggles more in the NIS 2 Directive, where precision and recall both decline, leading to the lowest F1 score among models for this document. These results reinforce the observation made in the per-document analysis: that NIS 2, due to its abstract legal formulations and fragmented answer locations, presents the greatest challenge for all models.

What these per-model, per-document results reveal is a deeper picture of specialization and fragility. DeepSeek's recall-maximizing strategy makes it highly suitable for information-seeking tasks that prioritize content coverage but raises concerns for use cases where answer compactness is essential. Gemma shows similar behavior but offers better performance in ENS, where it achieves an F1 of 0.7500—one of its best outcomes—suggesting it adapts better to semi-structured legal formats. Mistral's strong and consistent F1 scores across all three documents (all greater than 0.8235) suggest that it maintains the most robust performance in diverse normative environments. Its ability to deliver both relevant and focused answers across structured and abstract regulatory language makes it particularly well-suited for real-

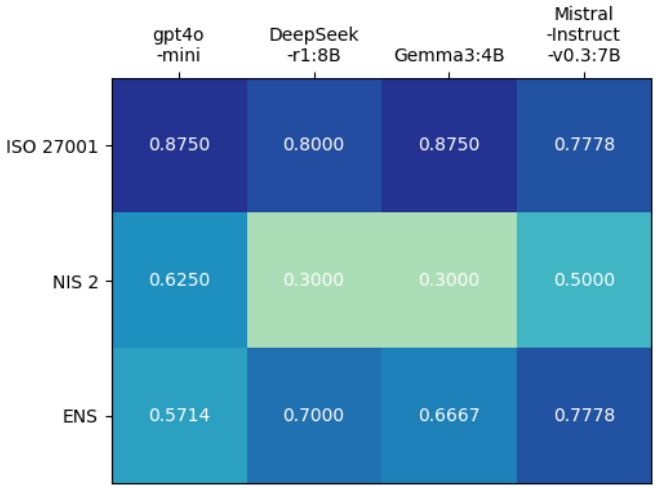


Fig. 3. Precision for each Model and Document

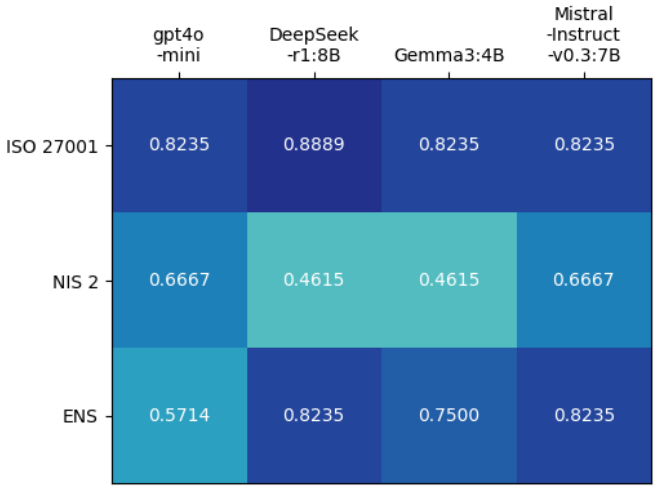


Fig. 5. f1-Score for each Model and Document

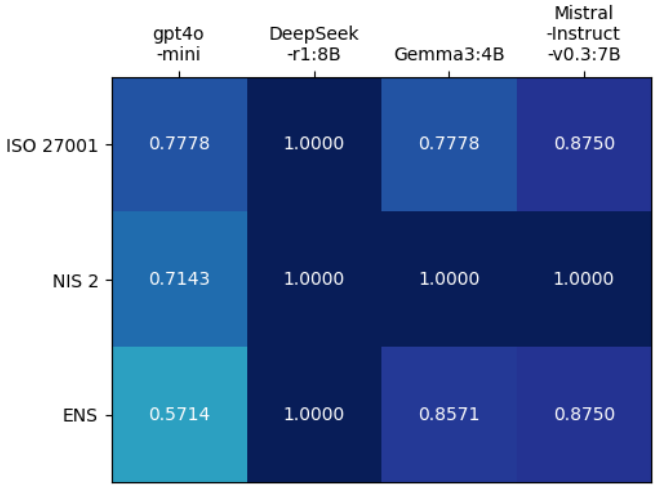


Fig. 4. Recall for each Model and Document

world applications requiring both breadth and specificity.

These findings suggest that, beyond overall averages, the interplay between model architecture and document structure plays a critical role in determining performance. Selecting an appropriate model for a given regulatory corpus may thus depend not only on the model's global metrics but also on its sensitivity to the stylistic and structural properties of the target text.

VI. Conclusions

In this paper,*CyberChatbot* has been presented. It is a AI-driven chatbot system based on RAG techniques specially designed to help Spanish-speaking users understand cybersecurity regulations.

The system uses official documents from ISO/IEC 27001, the ENS, and the NIS 2 directive to answer user questions in natural language. These documents were prepared and transformed to keep their structure, and the system was tested with different language models under the same conditions.

The experimental results showed that the proposed system can provide accurate and grounded answers in Spanish. Among the tested models, GPT-4o-mini, DeepSeek-r1:8B, and Mistral-Instruct-v0.3 performed better in precision, recall, and F1 score. This confirms that RAG-based systems are effective for helping users access and understand complex legal content. Thus, this solution makes cybersecurity regulations more accessible, especially for small organizations and professionals without deep legal or technical knowledge.

In the future, a key area involves scaling the system to operate over a larger and more diverse corpus of regulatory texts. This will be supported by refining the document ingestion pipeline with more sophisticated preprocessing and noise-reduction techniques to ensure high-quality, semantically coherent chunks. Agent-based mechanisms could be included to manage this issue.

At the model level, lightweight fine-tuning strategies could be explored. For instance, LoRA [23] and QLoRA [24] could be interesting to consider for adapting open-weight language models to the specific demands of the legal and regulatory context.

Acknowledgment

This research has been supported by grants from the Spanish Ministry of Science and Innovation, under the Knowledge Generation Projects program: XMIDAS (Ref: PID2021-122640OB-100), and the Public-Private Collaboration program: DICYME (Ref: CPP2021-009025).

Moreover, this initiative is carried out within the framework of the funds of the Recovery, Transformation, and Resilience Plan, financed by the European Union (Next Generation) - National Cybersecurity Institute (INCIBE) in the project C107/23 "Artificial Intelligence Applied to Cybersecurity in Critical Water and Sanitation Infrastructures".

References

[1] M. Masombuka, M. Grobler, and P. Duvenage, "Cybersecurity and local government: Imperative, challenges and priorities," in *ECCWS 2021 20th European Conference on Cyber Warfare and Security*, vol. 285. Academic Conferences Inter Ltd, 2021.

[2] International Organization for Standardization, "ISO/IEC 27001:2022 - Information security, cybersecurity and privacy protection – Information security management systems – Requirements," https://www.iso.org/standard/82875.html, 2022, accessed: 2025-04-07.

[3] European Parliament and of the Council, "Directive (EU) 2022/2555," https://eur-lex.europa.eu/eli/dir/2022/2555/oj/eng, 2022, accessed: 2025-04-15.

[4] Gobierno de España, "Real Decreto 311/2022, por el que se regula el Esquema Nacional de Seguridad," https://www.boe.es/eli/es/rd/2022/05/03/311, 2022, accessed: 2025-04-07.

[5] NIST, "Security and Privacy Controls for Information Systems and Organizations (SP 800-53 Rev. 5)," https://doi.org/10.6028/NIST.SP.800-53r5, National Institute of Standards and Technology, Tech. Rep., 2020, accessed: 2025-04-07.

[6] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," 2020. [Online]. Available: https://arxiv.org/abs/2005.11401

[7] R. Asyrofi, M. R. Dewi, M. I. Lutfhi, and P. Wibowo, "Systematic literature review langchain proposed," in *2023 International Electronics Symposium (IES)*. IEEE, 2023, pp. 533–537.

[8] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[9] F. Greco, G. Desolda, A. Esposito, A. Carelli *et al.*, "David versus goliath: Can machine learning detect llm-generated text? a case study in the detection of phishing emails," in *The Italian Conference on CyberSecurity*, 2024.

[10] S. Patel and V. K. Madisetti, "Phishguard: Integrating fine-tuned large language models (llms) into password management," *Journal of Information Security*, vol. 15, no. 4, pp. 474–493, 2024.

[11] C. Song, L. Ma, J. Zheng, J. Liao, H. Kuang, and L. Yang, "Audit-llm: Multi-agent collaboration for log-based insider threat detection," *arXiv preprint arXiv:2408.08902*, 2024.

[12] Advisera, "Experta – the first ai chatbot specialized for iso 27001, 9001, 14001, and other standards," Nov. 2024, accessed: 2025-04-07. [Online]. Available: https://advisera.com/experta/

[13] Better ISMS, "ISO 27001 Copilot – Your AI assistant for ISMS implementation," 2024, accessed: 2025-04-07. [Online]. Available: https://ismscopilot.com

[14] Botable.ai, "Botable – AI-powered chatbot for Compliance teams," 2024, accessed: 2025-04-07. [Online]. Available: https://www.botable.ai/departments/compliance

[15] Security Docs Guide Contributors, "Security Docs Guide Chatbot – Open-source assistant for cybersecurity compliance," 2024, gitHub repository, accessed: 2025-04-07. [Online]. Available: https://github.com/example/security-docs-chatbot

[16] C. Auer, M. Lysak, A. Nassar, M. Dolfi, N. Livathinos, P. Vagenas, C. B. Ramis, M. Omenetti, F. Lindlbauer, K. Dinkla, L. Mishra, Y. Kim, S. Gupta, R. T. de Lima, V. Weber, L. Morin, I. Meijer, V. Kuropiatnyk, and P. W. J. Staar, "Docling technical report," 2024. [Online]. Available: https://arxiv.org/abs/2408.09869

[17] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu, "Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation," 2024. [Online]. Available: https://arxiv.org/abs/2402.03216

[18] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, and H. Jégou, "The faiss library," 2024. [Online]. Available: https://arxiv.org/abs/2401.08281

[19] Z. Xie and C. Wu, "Mini-omni2: Towards open-source gpt-4o with vision, speech and duplex capabilities," *arXiv preprint arXiv:2410.11190*, 2024.

[20] DeepSeek-AI, D. Guo, D. Yang, H. Zhang, and J. S. ..., "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," 2025. [Online]. Available: https://arxiv.org/abs/2501.12948

[21] G. Team, A. Kamath, J. Ferret, S. Pathak, and N. V. ..., "Gemma 3 technical report," 2025. [Online]. Available: https://arxiv.org/abs/2503.19786

[22] Z. Gao, J. Deng, P. Reviriego, and S. Liu, "Operating conversational large language models (llms) in the presence of errors: The case of mistral-7b," 2024.

[23] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," 2021. [Online]. Available: https://arxiv.org/abs/2106.09685

[24] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," 2023. [Online]. Available: https://arxiv.org/abs/2305.14314

A Virtual Cybersecurity Department for Securing Digital Twins in Water Distribution Systems

1st Mohammadhossein Homaei
Media Engineering Group
University of Extremadura
Cáceres, Spain
mhomaein@alumnos.unex.es

2nd Agustin Di Bartolo
3rd Óscar Mogollón-Gutiérrez
Media Engineering Group
University of Extremadura
Cáceres, Spain
{adibartolo, oscarmg}@unex.es

4th Fernando Broncano Morgado,
5th Pablo García Rodríguez
Media Engineering Group
University of Extremadura
Cáceres, Spain
{fbroncano, pablogr}@unex.es

***Abstract*—Digital twins (DTs) help improve real-time monitoring and decision-making in water distribution systems. However, their connectivity makes them easy targets for cyberattacks such as scanning, denial-of-service (DoS), and unauthorized access. Small and medium-sized enterprises (SMEs) that manage these systems often do not have enough budget or staff to build strong cybersecurity teams. To solve this problem, we present a Virtual Cybersecurity Department (VCD), an affordable and automated framework designed for SMEs. The VCD uses open-source tools like Zabbix for real-time monitoring, Suricata for network intrusion detection, Fail2Ban to block repeated login attempts, and simple firewall settings. To improve threat detection, we also add a machine-learning-based IDS trained on the OD-IDS2022 dataset using an improved ensemble model. This model detects cyber threats such as brute-force attacks, remote code execution (RCE), and network flooding, with 92% accuracy and fewer false alarms. Our solution gives SMEs a practical and efficient way to secure water systems using low-cost and easy-to-manage tools.**

***Index Terms*—Digital Twins, Cybersecurity, Intrusion Detection System, Machine Learning, Zabbix, Water Distribution, SMEs**

I. Introduction

As water distribution systems become increasingly connected, they face growing cybersecurity risks [1]–[6]. Integrating information technology (IT) and operational technology (OT) has significantly improved efficiency and real-time monitoring, but has also introduced new vulnerabilities. DT technology, which provides virtual replicas of physical systems, further enhances these capabilities by improving operational visibility, predictive maintenance, and decision-making [7], [8]. However, increased connectivity and intelligence in DTs expand their attack surface, making them vulnerable not only to data leaks but also to threats that could impact public health and infrastructure safety. Cyberattacks such as unauthorized access, data manipulation, or DoS could result in severe incidents like water contamination or system disruptions [9]. These attacks can go undetected for long periods, especially in systems without active monitoring. As the number of smart sensors and connected devices grows, the complexity of protecting the infrastructure increases. Thus, robust, automated cybersecurity measures are crucial.

SMEs, which often manage water distribution networks, have serious challenges in cybersecurity because they usually do not have enough money or trained IT staff. Traditional security systems are expensive and need expert teams, so they are not good options for small organizations. To solve this, we propose a VCD, a low-cost and easy-to-use system built with open-source tools. The main tool in the VCD is Zabbix, which gives real-time system monitoring, alerting, and data visualization [10], [11]. It helps detect technical problems and possible cyber threats in the digital twin environment. To improve detection, we also added a machine-learning-based IDS trained on the OD-IDS2022 dataset. This system can find different types of cyberattacks, such as scanning, brute-force, RCE, and DoS attacks. By combining simple monitoring with advanced machine learning, our framework gives SMEs an effective and affordable way to protect their water systems.

Unlike many existing frameworks, our VCD uniquely combines traditional open-source tools with a customized, explainable machine learning model, all optimized for low-resource environments typical in SME water utilities.

The motivation for this work comes from real needs in the field. Many small and rural water utilities want to use digital twin systems, but they are not ready to face growing cybersecurity risks. Most existing solutions are made for large companies and need expensive tools or professional IT teams. SMEs cannot afford these systems and are left with weak protection. Also, new cyberattacks are becoming smarter and harder to detect with old methods. Our goal is to offer a practical and affordable solution that helps SMEs protect their water infrastructure using tools they can manage themselves. The proposed Virtual Cybersecurity Department gives them a way to use open-source software, automate responses, and improve detection with machine learning—without needing large investments or complex systems.

The remainder of the paper is organized as follows: Section II reviews existing research in cybersecurity for DT-based water systems. Section III introduces our proposed VCD framework, including system architecture, communication flow, cybersecurity integration, and ML-based IDS. Section IV presents the experimental evaluation, including results from the Zabbix-based monitoring and the IDS model performance. Finally, Section V concludes the paper and outlines directions for future work.

TABLE I
RECENT WORK ON CYBERSECURITY IN DT-ENABLED WATER DISTRIBUTION SYSTEMS (POST-2020)

Ref.	Focus, Challenges, and Tech. & Eval.		
	Focus	***Challenges***	***Tech. & Eval.***
Zhang *et al.* 2021 [12]	Attack detection in DT water systems	Distinguish anomalies from normal ops; Integrate IT/OT data	ML anomaly detection; IoT integration; Simulation testbed
Liu *et al.* 2022 [13]	Secure smart water DTs	Ensuring secure communication	Cryptographic protocols; Anomaly-based IDS; Emulated DT with intrusions
Qi *et al.* 2022 [14]	Risk assessment in DT networks	Prioritizing vulnerabilities in distributed systems	Sensor fusion; Statistical threat scoring; Risk evaluation scenario analysis
Kumar *et al.* 2023 [15]	Mitigate attacks via anomaly+blockchain	Data tampering, traceability	Blockchain for data integrity; ML detection; Experimental deployment
Lin *et al.* 2023 [16]	IDS using DT correlation (hydraulic/network)	Detect stealthy attacks in normal ops	Hybrid IDS correlating physical & network metrics; Lab-scale DT with synthetic attacks

II. RELATED WORK

A. Cybersecurity in DTs for Water Systems

In recent years, DT technology has become more common in water distribution systems due to its ability to provide real-time monitoring, predictive analytics, and decision support. However, this increased connectivity has also introduced new cybersecurity challenges. DTs, by design, connect multiple physical and digital components, which increases the attack surface for potential cyber threats.

Zhang *et al.* [12] presented a machine-learning-based intrusion detection framework for DT-enabled water systems, integrating IoT sensors to detect anomalies in physical and cyber operations. Liu *et al.* [13] proposed a secure DT architecture using encryption protocols and anomaly-based IDS to protect communication flows between devices and the cloud. Homaei *et al.* [1], [2] also highlighted the dual role of DTs as both monitoring tools and high-risk targets for attacks, especially in rural water networks.

These studies show that while DTs improve operations, they also require new security solutions that go beyond traditional IT protections.

B. Challenges in DT-based Water Infrastructure

Cybersecurity in water distribution systems faces multiple technical and operational challenges, particularly when DTs are integrated:

- Anomaly Detection: DT systems rely on normal behavior patterns to function correctly. However, cyberattacks often mimic legitimate fluctuations (e.g., consumption peaks), making detection difficult without advanced ML techniques.
- Scalability and Performance: Real-time monitoring and analysis require high processing power and efficient algorithms, especially as the number of IoT sensors increases.
- Legacy and Modern System Integration: Many utilities still use legacy systems that are not easily compatible with modern IoT devices or secure communication protocols, creating interoperability issues.
- Network Communication Risks: Protocols used in DTs are sometimes unencrypted or misconfigured, exposing them to packet sniffing, spoofing, or DoS attacks.
- Limited Resources in SMEs: Most SMEs lack the IT staff, funding, or training to maintain enterprise-level cybersecurity systems, leaving them especially vulnerable.
- Public Safety and Reliability: Failures in cyber-protected DTs could lead to water shortages, contamination, or service disruptions, affecting entire communities.

These issues make it clear that new frameworks should be lightweight, scalable, and capable of operating in low-resource environments.

C. Emerging Solutions and Gaps

Recent research has introduced several approaches to improve cybersecurity in DT-enabled water networks. Qi *et al.* [14] introduced a risk assessment method using sensor fusion and statistical analysis to identify vulnerable components. Kumar *et al.* [15] proposed combining blockchain with anomaly detection to increase data traceability and prevent tampering. Lin *et al.* [16] focused on hybrid intrusion detection systems that analyze both physical process data and network logs to detect stealth attacks.

Although these methods show progress, they often rely on complex systems or high-performance resources, which may not be suitable for SMEs.

D. Positioning of This Work

In contrast to prior works that require extensive infrastructure or expert personnel, our proposed VCD offers a practical alternative for small and medium-sized enterprises. The VCD uses a combination of lightweight, open-source tools—Zabbix, Suricata, and Fail2Ban—alongside a machine-learning-based IDS trained on the OD-IDS2022 dataset.

Unlike many traditional systems that depend solely on signature-based detection or manual log review, our model integrates real-time monitoring with automated responses and a trained ensemble ML model. This hybrid approach improves detection of advanced threats such as brute-force, RCE, and DoS attacks, making it well-suited for decentralized water systems with limited resources. It also reduces the need for continuous human supervision and simplifies system maintenance, allowing operators to focus on operational tasks rather than complex cybersecurity management.

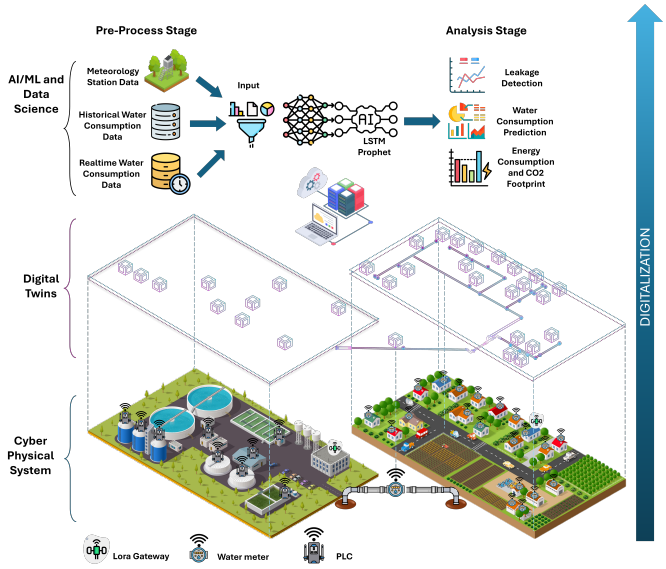


Fig. 1. DT platform in the WDS [17]

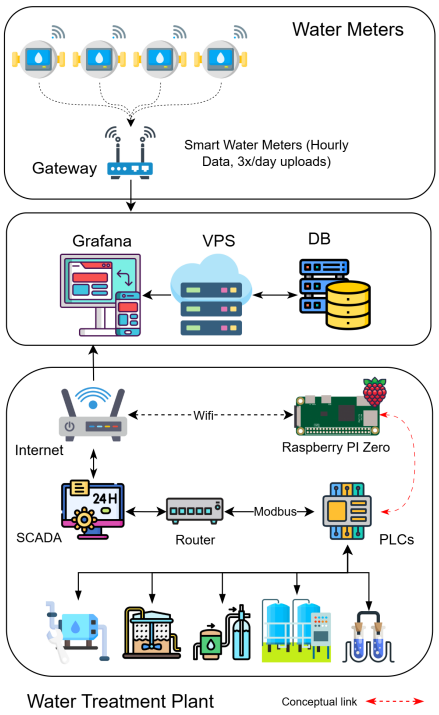


Fig. 2. Deployment of Zabbix proxies on Raspberry Pi devices for real-time data collection from IoT meters and SCADA systems.

III. Proposed Framework

This section describes the structure and components of the proposed VCD, a cost-effective monitoring framework for DTs in WDS. The system is designed to help SMEs enhance their operational security through automated, open-source tools. The framework includes four main components: the DT system overview, system architecture and communication flow, cybersecurity integration using Zabbix and Suricata, and a machine learning-based IDS.

A. DT System Overview

The VCD is built on a Digital Twin platform that integrates real-time data collection, AI-driven analytics, and secure communication. It consists of three main layers: cyber-physical systems (CPS), data management, and predictive analytics.

The CPS layer includes sensors, PLCs, and IoT water meters deployed in water treatment facilities and distribution pipelines. These devices collect environmental, operational, and consumption data. The data is transmitted securely using technologies such as LoRaWAN, VPN, and SSH. AI/ML models—including LSTM, Prophet, and LightGBM—are used for water usage forecasting, leakage detection, and energy monitoring. Additionally, GIS tools support spatial analysis and map-based monitoring (Figure 1) [17].

This architecture is designed for rural and small-scale water utilities but is scalable for larger infrastructures. It provides enhanced operational control, cost efficiency, and resource optimization.

B. System Architecture and Communication Flow

The system consists of three key components: edge nodes, secure communication channels, and a central server.

- Edge nodes include Raspberry Pi devices equipped with Zabbix proxies, IoT meters, SCADA units, and PLCs. These nodes are strategically placed at water plants and administrative locations to ensure complete visibility (Figure 2).
- Secure communication is established using VPN tunnels, SSH protocols, and LoRaWAN networks. This ensures that data from the field devices reaches the central server with integrity and confidentiality. Zabbix continuously monitors the stability and quality of these connections.
- The central server, hosted on a Virtual Private Server (VPS), aggregates all incoming data. It runs Zabbix for real-time monitoring, Suricata for intrusion detection, and Fail2Ban for automated IP blocking. The server can optionally connect to cloud platforms like AWS or Azure for data storage and computational scalability.

Figure 3 presents the VCD architecture, highlighting the placement of Zabbix and the ML-based IDS modules.

C. Cybersecurity Integration with Zabbix and Suricata

The core of the cybersecurity layer is Zabbix, which provides data collection, visualization, and alerting functionalities. It monitors metrics such as network traffic, CPU load, memory usage, and failed login attempts. Zabbix is integrated with Suricata, an open-source IDS that inspects network packets and detects threats like port scanning, brute-force logins, and unusual data flows. Suricata's alerts are visualized in the Zabbix dashboard. Fail2Ban complements the system by monitoring authentication logs. It automatically bans IP addresses that exceed a defined number of failed login attempts. This combination of tools ensures multi-layered protection against a

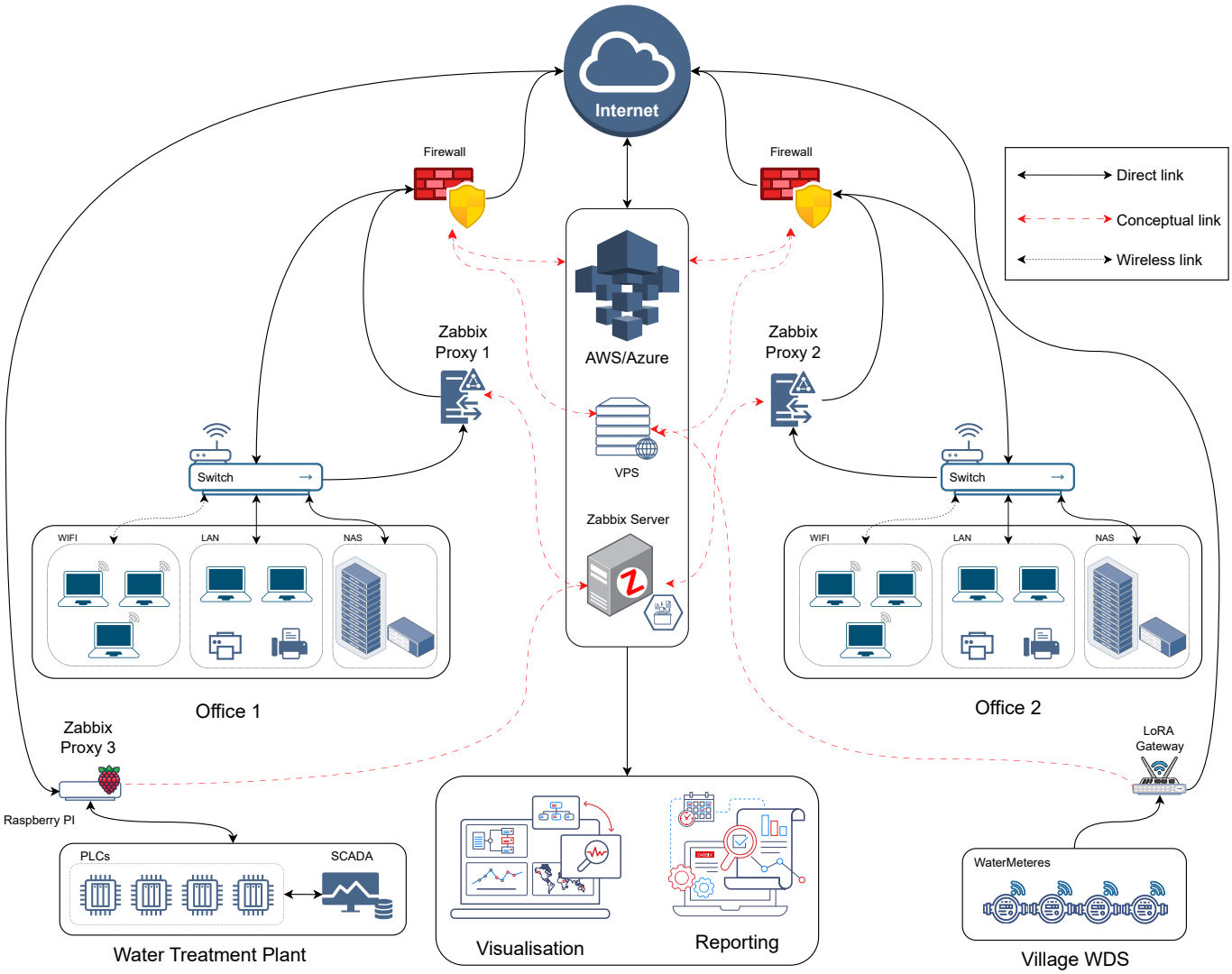


Fig. 3. VCD architecture with Zabbix and ML-based IDS for DT-enabled SME water systems

wide range of attacks while remaining lightweight and suitable for resource-constrained environments.

D. AI/ML-Based Intrusion Detection System

As part of the proposed framework, we developed a machine learning-based IDS to improve the detection of cyber threats in smart water networks. This IDS is trained on the OD-IDS2022 dataset, which provides 1,031,916 labeled samples [18]. Each sample contains 82 features representing flow-based network data, including IP addresses, port numbers, protocol types, packet lengths, time durations, and flag behaviors. These records include normal traffic and 29 attack types such as DoS, brute force, SQL injection, RCE, hijacking, and reconnaissance. To simplify classification and reduce overfitting, we grouped the 29 attack classes into seven general categories, listed in Table II. This grouping keeps the detection meaningful while making the machine learning models easier to train and evaluate.

TABLE II
7-GROUP ATTACK CATEGORIZATION

Group	Includes
BENIGN	BENIGN
DOS	DoS Hulk, Slowhttptest, GoldenEye, Slowloris, DDoS-*
BRUTEFORCE	Bruteforce-Web, Bruteforce-XSS, FTP/SSH-Patator, Web Brute Force
INJECTION	SQL/LDAP/SIP Injection, Web SQL Injection
HIJACKING	MITM, Hijacking
RCE	RFI, Exploit, Cmd Injection, Upload, Backdoor
OTHER	Infiltration, Bot, PortScan, Web XSS

We proposed and implemented five machine learning models as part of the IDS component. All models use the same preprocessing pipeline: label encoding, numerical feature extraction, mutual information for feature selection, data normalization, and oversampling with SMOTE to balance class distribution.

1) Random Forest Classifier: The Random Forest (RF) model builds many decision trees from random subsets of the training data. Each tree gives a class prediction, and the final result is selected by majority voting. This is expressed in

Equation 1.

$$\hat{y} = \text{mode}\left(h_1(x), h_2(x), \ldots, h_T(x)\right) \tag{1}$$

where $h_t(x)$ is the prediction of tree t, and T is the total number of trees.

2) Tuned LightGBM Classifier: LightGBM is a gradient boosting algorithm that builds trees sequentially to minimize prediction errors. It grows trees leaf-wise and uses a loss function with regularization, as shown in Equation 2.

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} L(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{2}$$

Here, L is the loss function, $\hat{y}_i^{(t-1)}$ is the previous prediction, f_t is the new decision tree, and Ω is the regularization term.

3) Improved Ensemble Model (v1): This model combines three base classifiers—Random Forest, LightGBM, and a Multi-layer Perceptron (MLP)—into a soft voting ensemble. It averages the class probabilities from each model and selects the class with the highest average score, as shown in Equation 3.

$$\hat{y} = \arg\max_c \left(\frac{1}{M} \sum_{m=1}^{M} P_m(c \mid x) \right) \tag{3}$$

where $P_m(c \mid x)$ is the probability of class c predicted by model m, and M is the number of models.

4) Weighted Ensemble with Feature Engineering: This model improves ensemble voting by assigning custom weights to each classifier and using new engineered features like packet length ratios and size variations. The prediction formula with weights is given in Equation 4.

$$\hat{y} = \arg\max_c \left(\sum_{m=1}^{M} w_m \cdot P_m(c \mid x) \right) \tag{4}$$

where w_m is the weight assigned to model m, and $\sum w_m = 1$.

5) Improved Ensemble (v2): The final and most optimized model uses the same weighted voting as in Equation 4, but with improved components. These include:

- A deeper MLP with 3 hidden layers (256, 128, 64) and ReLU activation
- A tuned LightGBM with max depth = 10, 64 leaves, and learning rate = 0.05
- A larger Random Forest with 150 trees and class-balanced weighting

The ensemble weights are selected based on validation scores to ensure balanced detection across all classes, especially minority attacks like RCE and Hijacking.

Note: To improve model transparency, we use SHAP (SHapley Additive exPlanations), a method from cooperative game theory that attributes prediction changes to individual features. The SHAP value for a feature i is calculated using:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f(S \cup \{i\}) - f(S) \right] \tag{5}$$

Here, F is the full feature set, S is a subset of features excluding i, and f is the prediction function. SHAP values explain how much each feature contributes to the final prediction, helping operators understand the decision process of the IDS.

In summary, the AI-based IDS module strengthens the proposed framework by enabling real-time detection of various cyber threats using interpretable and resource-efficient machine learning models. In the following section, we present the experimental evaluation and performance results of the IDS models, along with the integration of Zabbix for continuous monitoring in the digital twin environment.

IV. Experimental Evaluation and Model Performance

This section presents the experimental evaluation of the proposed VCD for water systems. The validation includes two parts: real-time monitoring results using Zabbix, and the performance of machine learning models for intrusion detection.

A. Monitoring Setup and Attack Simulation

The VCD was tested in a hybrid digital twin setup. Zabbix server was installed on a VPS, and several Raspberry Pi devices were installed in field locations like water plants and offices. These Raspberry Pis worked as Zabbix proxies and collected logs from IoT meters, PLCs, and SCADA systems.

To test the system, three types of cyberattacks were performed:

- *Nmap Scan (Reconnaissance)*: A stealth scan was launched using Nmap to find open ports. Suricata detected this scan and sent alerts to Zabbix. The traffic pattern showed abnormal packet behavior (Figure 4).
- *Brute Force (Hydra + SSH)*: An SSH brute-force attack was simulated using Hydra. Zabbix recorded many failed logins and increased CPU usage. Suricata also detected frequent access to port 22. Fail2Ban blocked the attacker's IP after too many failed attempts, as shown in Figures(5, 6).
- *DoS (hping3)*: A SYN flood attack was done using hping3. It caused high CPU and memory usage. Zabbix showed this unusual behavior and created alerts, even when the logs were not clear.

These tests showed that the system can detect and respond to real cyberattacks using simple and open-source tools.

B. Monitoring Indicators

Several indicators were collected from Zabbix to check the system behavior:

- *CPU and Memory Usage*: These increased during DoS attack and helped to detect it (Figure 7).

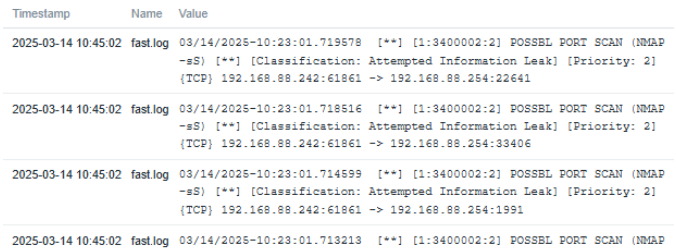


Fig. 4. logging attempt to the servers

Fig. 5. Fail2Ban logs showing IP bans triggered by repeated failed SSH login attempts

- *Network Traffic (Upload/Download)*: Abnormal traffic helped detect Nmap and DoS attacks (Figure 8).
- *Dropped and Malformed Packets*: These increased during the flood attack.
- *Failed Login Attempts*: Zabbix tracked this for brute force detection, and Fail2Ban blocked the IP.
- *Suricata Alerts*: Number of alerts helped show which attack was happening.
- *Alert Time*: The system created alerts in a few seconds after the attack started.

C. Machine Learning IDS Evaluation

Besides traditional detection, five machine learning models were tested using the OD-IDS2022 dataset. The goal was to

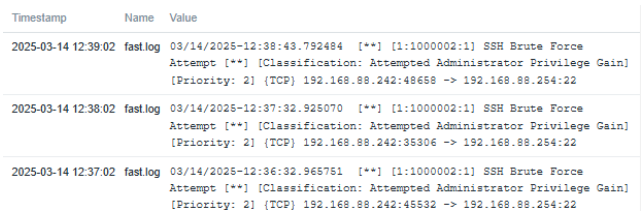


Fig. 6. Suricata alerts for SSH brute-force attack attempts showing repeated unauthorized access to port 22

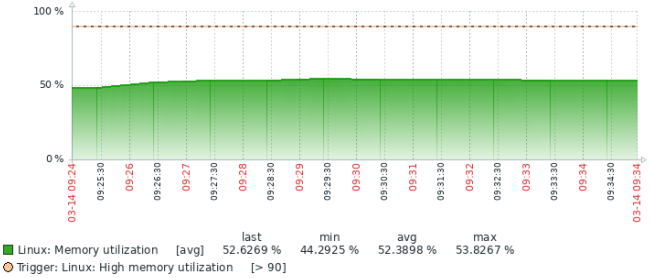


Fig. 7. Memory usage monitoring under DDoS Attack

Fig. 8. Network monitoring under DDoS Attack

classify seven types of network traffic, including attacks like RCE, hijacking, and injection.

Table III shows the performance of each model. The best model was the improved ensemble (v2), which used LightGBM, Random Forest, and MLP together.

TABLE III
COMPARISON OF IDS MODELS FOR 7-CLASS CATEGORIZATION (OD-IDS2022 DATASET)

Model	Acc.	Macro F1	RCE F1	HIJACK Rec.	Explainable	Gran.
Random Forest	77.0%	0.47	0.37	0.54	✓ SHAP	30+
LightGBM (Tuned)	82.2%	0.714	0.526	0.609	– (addable)	7
Improved Ens. (v1)	80.4%	0.6645	0.55	0.73	✓ SHAP	7
Weighted Ens. + FE	80.2%	0.66	0.54	0.76	✓ SHAP	7
Improved Ens. (v2)	92.0%	0.88	0.86	0.87	✓ SHAP	7

Table IV shows the full report for the best model. It gives good results in all classes, including small ones like injection. Figure 9 shows the confusion matrix.

TABLE IV
ENSEMBLE MODEL CLASSIFICATION REPORT

Class	Precision	Recall	F1-score	Support
BENIGN	0.91	0.93	0.92	2024
BRUTEFORCE	0.87	0.85	0.86	2377
DOS	0.96	0.97	0.96	6463
HIJACKING	0.84	0.87	0.85	2475
INJECTION	0.82	0.78	0.80	220
OTHER	0.92	0.93	0.93	14629
RCE	0.85	0.88	0.86	1812
Accuracy			0.92	30000
Macro Avg	0.88	0.89	0.88	30000
Weighted Avg	0.92	0.92	0.92	30000

This experiment confirms that the proposed VCD can detect and respond to cyberattacks in real time using both rule-based

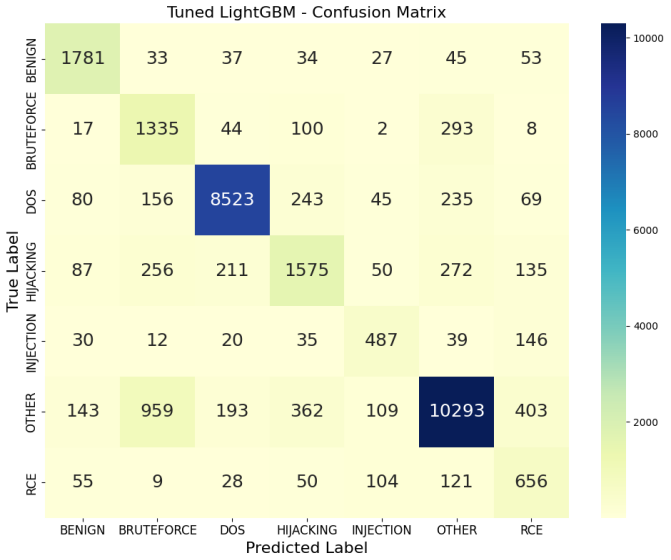


Fig. 9. Confusion matrix showing class-wise prediction performance across 7 traffic categories.

and AI-based tools. It works well even in small water systems with low-cost hardware.

V. Conclusion and Future Work

This study proposed a VCD to improve the cybersecurity of DTs used in water networks, with a focus on SMEs. The solution combines free tools: Zabbix for live monitoring, Suricata as an IDS, and Fail2Ban to block repeated login attempts. Zabbix proxies were installed on Raspberry Pi units to collect data from SCADA, PLCs, and IoT sensors. We tested the system with simulated attacks (port scanning, brute-force on SSH, and DoS), and it responded correctly with alerts, log collection, and IP blocking.

The IDS part was developed using the OD-IDS2022 dataset (over one million records with 29 classes). We simplified the task by grouping the classes into 7 attack types. We tested five ML models, and the final version (v2) used a combination of RF, LGBM, and MLP with SHAP for explainability. This model gave the best results for detecting attacks like RCE, hijacking, and injection. The full framework is low-cost, supports real-time detection, and works well for small organizations without advanced computing systems.

For future work, we will explore LLMs to improve detection accuracy, reduce false positives, and classify threats more precisely. We also plan to integrate blockchain to protect data integrity and support trusted operations. These upgrades aim to create smarter and more secure water management systems.

Acknowledgment

This initiative is carried out within the framework of the funds from the Recovery, Transformation, and Resilience Plan, financed by the European Union (Next Generation) – National Institute of Cybersecurity (INCIBE), as part of project C107/23: "Artificial Intelligence Applied to Cybersecurity in Critical Water and Sanitation Infrastructures."

References

[1] M. Homaei, O. Mogollón-Gutiérrez, J. C. Sancho, M. Ávila, and A. Caro, "A review of digital twins and their application in cybersecurity based on artificial intelligence," *Artificial Intelligence Review*, vol. 57, no. 8, jul 2024.

[2] M. Homaei, A. C. Lindo, J. C. S. N. nez, O. M. Gutiérrez, and J. A. Díaz, "The role of artificial intelligence in digital twin's cybersecurity," in *XVII Reunión Española Sobre Criptología y Seguridad de La Información (RECSI)*, vol. 265, 2022, p. 133.

[3] A. Abbasi, F. Zaidi, and O. F. Rana, "A survey on cybersecurity in critical infrastructures: Approaches, challenges, and future directions," *ACM Computing Surveys*, vol. 54, no. 7, pp. 1–36, 2021.

[4] J. Smith and A. Brown, "A review of cyber threats and security solutions for water distribution networks," *Journal of Industrial Information Integration*, vol. 15, pp. 100–110, 2019.

[5] S. Yu and L. Liu, "Survey on cyber-physical system security in water sector: Threats and defense strategies," *IEEE Access*, vol. 9, pp. 78 765–78 779, 2021.

[6] A. Brown, C. Wu, and D. Wilson, "Anomaly detection in critical infrastructures: A survey of methods and challenges," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 3, pp. 177–198, 2022.

[7] F. Tao and M. Zhang, "Digital twin shop-floor: A new shop-floor paradigm towards smart manufacturing," *IEEE Access*, vol. 5, pp. 20 418–20 427, 2018.

[8] A. Fuller, Z. Fan, C. Day, and C. Barlow, "Digital twin: Enabling technologies, challenges and open research," *IEEE Access*, vol. 8, pp. 108 952–108 971, 2020.

[9] A. Mirchi and K. Madani, "Water resources cyber-physical security: A review and future research directions," *Water*, vol. 12, no. 6, p. 1602, 2020.

[10] Z. SIA, "Zabbix: Open source monitoring solution," https://www.zabbix.com, 2025, accessed: April 15, 2025.

[11] V. Kandasamy and M. Shankaran, "Survey on open-source ids and monitoring solutions for critical infrastructures," *International Journal of Network Security*, vol. 23, no. 1, pp. 45–58, 2021.

[12] Z. Zhang, X. Chen, G. Zhao, and W. Gao, "A cyber-physical attack detection framework for digital twin-based water distribution systems," *IEEE Internet of Things Journal*, vol. 8, no. 9, pp. 7650–7660, 2021.

[13] D. Liu, M. Zhong, Y. Fu, and Y. Li, "Cybersecurity for digital twin-based smart water management systems," *Journal of Water Resources Planning and Management*, vol. 148, no. 4, p. 04022005, 2022.

[14] S. Qi, K. Wang, and H. Zhuang, "Cybersecurity risk assessment in digital twin-enabled smart water networks," *Sensors*, vol. 22, no. 18, p. 6905, 2022.

[15] S. Kumar, Y. Wu, and J. Li, "Mitigating cyber-attacks in digital twin environments of water distribution networks using anomaly detection and blockchain," *IEEE Transactions on Industrial Informatics*, 2023, early Access.

[16] Y. Lin, R. Deng, and H. Chen, "Securing cyber-physical water distribution infrastructures: A digital twin-based intrusion detection approach," *Journal of Network and Computer Applications*, vol. 225, p. 103525, 2023.

[17] M. Homaei, A. J. Di Bartolo, M. Ávila, O. Mogollón-Gutiérrez, and A. Caro, "Digital transformation in the water distribution system based on the digital twins concept," 2024.

[18] N. D. Patel, B. M. Mehtre, and R. Wankar, "Od-ids2022: generating a new offensive defensive intrusion detection dataset for machine learning-based attack classification," *International Journal of Information Technology*, vol. 15, no. 8, p. 4349–4363, Sep. 2023.

A Hands-On Learning Platform for CVE Understanding

Afonso Vitório and João Paulo Barraca
Instituto de Telecomunicações, DETI, Universidade de Aveiro, Portugal

***Abstract*—Cybersecurity is of uttermost importance for our society structure and digital infrastructures, and mechanisms such as CVE, have been created for cybersecurity professionals to better enumerate and communicate regarding existing public vulnerabilities. This very widely used technology was created in the finals of the '90s, and did not consider the tens of thousands of vulnerabilities found each year. Parallel to this, there has been a shift in cybersecurity learning, with more and more professionals and students learning about cybersecurity in a hands-on or gamified way. In this work, we target this trend, and aim to help professionals and students to better learn about CVEs by proposing a platform that contains a search engine for CVEs, showcasing information like PoCs available, and vulnerable VMs to allow users to learn about each CVE in a hands-on way.**

I. Introduction

In the current day and age, our society relies on the digital infrastructure, which is susceptible to a wide range of attacks. When malicious actions are taken by threat actors, the impacts can be catastrophic outside the information technology scope, also impacting the physical world, individuals and the society. The impacts can be so severe that many experts consider them the most relevant enemy of the modern economy. And while, the consequences of cybercrime are severe, cybercrime is on the rise year after year [1] [2], as the definition of crime is frequently a matter of jurisdiction.

Organizations rely heavily on complex technological stacks, that often include outdated technologies. Older technologies tend to have more security flaws, are better known by attackers, but constitute an ever increasing burden that blue teams must follow in order to keep an organization secure.

The most popular and widely used standard to catalog known security flaws is developed around Common Vulnerabilities and Exposures [3]. From a defenders perspective, teams a great need to learn about Common Vulnerabilities and Exposuress (CVEs) present in critical organizational assets, detect if the CVEs are being actively explored, identify potential indicators and then apply a risk-based remediation [4]. This is difficult for well-known software, and is a daunting task for a blue team defending the wide surface of current organizations.

Platforms specialized in hands-on and gamified learning for cybersecurity such as HackTheBox[1] and TryHackMe[2] have a massive user base, with over 2 million users on each, highlighting the demand for this type of learning approach. They also empower professionals to prioritize practical skills, as they want people that can apply specific competences and knowledge, and not just recite theory.

After this introduction, this paper is structured as follows. Section II presents relevant works related to our proposal. Section III describes our solution. Section IV describes the evaluation and highlights the main results. Section V presents our main conclusions and future work.

II. Related Work

According to the literature, there is a change happening in education, specifically in cybersecurity learning. Back in the day, education was just text-based learning. Today, it is leaning more towards hands-on labs, simulated environments, gamified learning, and the usage of tools such as Artificial Intelligence (AI) and Machine Learning (ML). This is proven to be a faster, and better way to learn about several topics [5, 6].

However, most people that want to learn about a CVE, have to rely heavily on text-based learning to learn about a CVE. If they want to get a deeper knowledge about CVEs, they have to build a custom lab, they have to get the vulnerable software in the specific version manually and build the testing environment themselves. This proves to be a very ineffective and time consuming process to learn about a CVE [7].

Furthermore, given the increase in the number of CVEs discovered every year, it is very hard for professionals to keep up, without better tools to learn about CVEs.

CVE databases such as National Vulnerability Database (NVD)[3], MITRE[4] and CVEDetails[5] provide text-based information but lack other types of information such as Proof of Concept (PoC) code, vulnerable Virtual Machines (VMs) and ways to experiment directly with the vulnerabilities. We specifically aim to fill this gap.

Another of the main hurdles in CVE learning is the problem of scattered resources. When a security professional or student tries to truly understand a vulnerability, they cannot just pull up one clean source and get everything they need. Instead, they're forced to hunt across many websites, if you want a description you have to go to a website, but if you want a working exploit, you have to dig through many sites and random GitHub repository, or blog posts written by independent researchers [8]. Sometimes the PoC code is buried in obscure forums or

[1]https://www.hackthebox.com
[2]https://www.tryhackme.com
[3]https://nvd.nist.gov
[4]https://www.mitre.org
[5]https://cvedetails.com

outdated write-ups, and even when you find it, it might not be compatible with modern systems.

In addition to the scattered nature of the available resources, there is also a lack of practical environments that are easily accessible for direct experimentation. Existing learning platforms often focus on theoretical content or general cybersecurity challenges rather than specifically addressing individual CVEs with realistic vulnerable environments. As a result, learners miss out on the essential hands-on experience that is critical for fully understanding how vulnerabilities operate in real-world systems [9]. Without streamlined access to ready-to-use, vulnerability-specific virtual machines and corresponding exploitation tools, the learning process remains fragmented, inefficient, and far removed from real-world attack scenarios.

III. A SOLUTION FOR IMPROVING CVE LEARNING

Our solution is based on a Web Application, that uses a hands-on learning approach to teach the user about a CVE. This platform acts not only as a sandbox, but also as a search engine, in which the user can search for a technology or a specific CVE and get the typical information associated to a CVE, plus the Exploit Prediction Scoring System (EPSS) score, the public PoC code if any is available and public docker images vulnerable to this CVE, ready to deploy in the Web Application.

In this web application the user is able to deploy vulnerable docker images and explore the vulnerability via a attacker VM that is accessible inside the Web Application via Secure Shell (SSH) or Virtual Network Computing (VNC). The user is also able to access the victim machine using SSH via the web application and have access to all sorts of information such as logs or a network capture of the traffic to the docker container.

A. Solution Architecture and Implementation

Our solution consists of a web application that aims to serve as a search engine and experimentatoin playform for CVEs. It retrieves generic information about a CVE and all the available public exploits from both GitHub[6] and ExploitDB[7]. Then it will return available vulnerable docker images.

This web application should also be able to create a test environment for the user to try and test the available PoCs. From the web app, the user should be able to deploy an attacker machine that will serve as the machine from which the exploits should be launched. It should also be able to deploy a victim VM on which the vulnerable docker images will be deployed. Inside the page with the results for a CVE, the user should be able to deploy a vulnerable docker image if it exists.

The architecture is depicted in Figure 1, and is composed by key modules that we will describe in the following paragraphs.

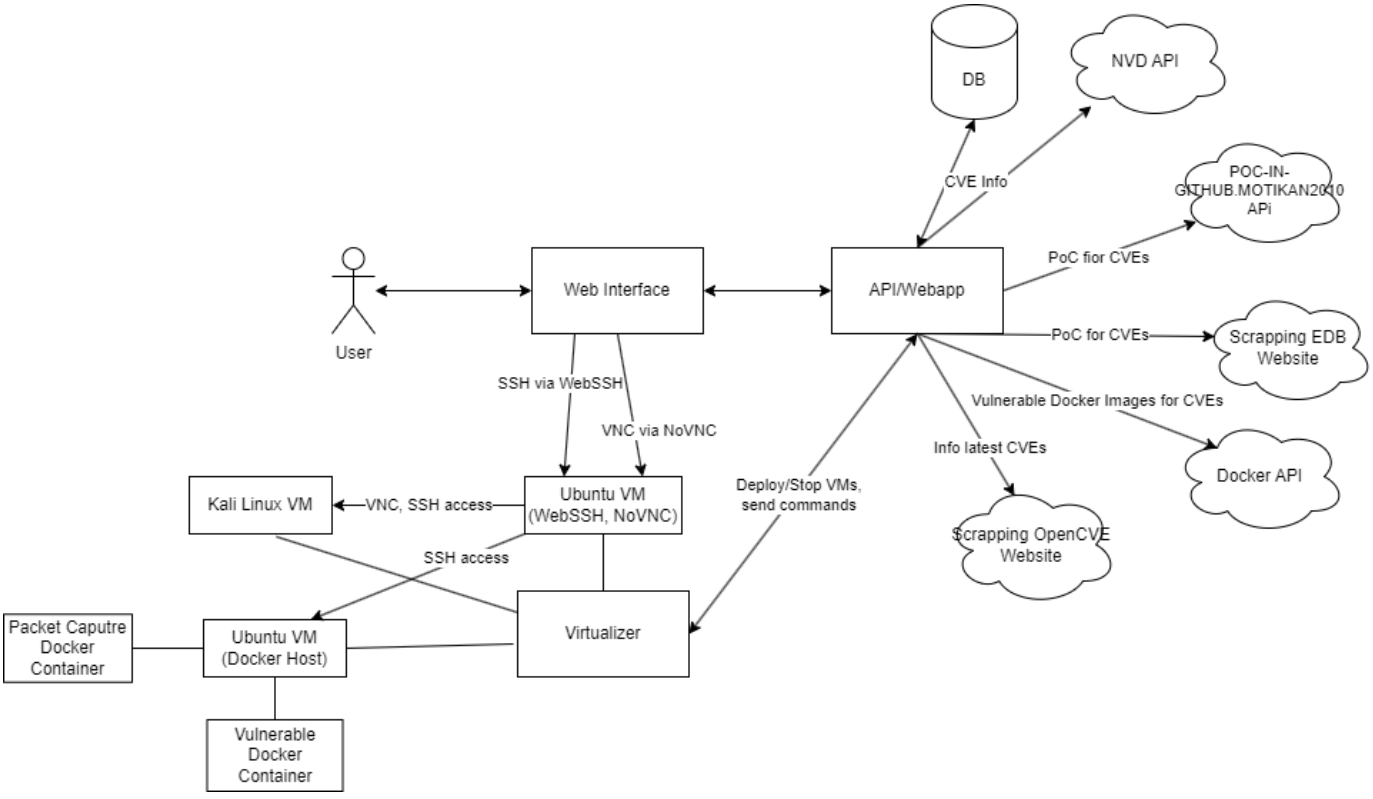


Figure 1. Architectural diagram of the the CVEBox application.

[6]https://www.github.com
[7]https://www.exploit-db.com

These components are the Web Interface, which is the component that the user is going to use to interact with the web application, get the information it needs, and perform the actions that need to be performed. The API/Webapp is the part responsible for gathering the information needed for the application to work and serve it to the Web Interface and is also the component responsible for interacting and performing actions to the VM. There is also the virtualizer and its multiple VMs, which are the machines responsible for the user to be able to dynamically test and interact with the PoC code for a CVE.

There is also a VM which is going to serve as the entry point to this network of VMs, it will be a server for tools that allow remote access inside of the network, which allows the user to have SSH and VNC access in the browser. There is also another VM, that will serve as the host for the vulnerable docker containers, and a VM that will serve as the attacker machine, from which the exploits will be launched.

The front-end interface on which the user can search for a specific CVEs or for a technology and get the ones associated.

In this interface, the user can select a CVE and then can get more detailed information about it, such as description, the PoC exploit code found from 2 websites, Exploit-DB and Github. This interface also provides the user with vulnerable docker images from docker hub and from Vulhub Github repository[8].

Figure 2. Homepage of the web application.

This page also contains the virtualization control on which the user is able to control, like start and stop, and access via either SSH or VNC, an attacker virtual machine, all from the web application. He can also control and access, in a similar fashion, the victim machine. This machine is responsible for hosting the docker containers deployed from the docker images found on this page.

Inside the page, for the CVEs we have six areas. The first area contains information regarding the CVEs, the description of the vulnerability. The **cvss!** (**cvss!**) score, which rates from 0 to 10 the importance of this vulnerability. The severity of this vulnerability, can range from Low, to Medium, to High to Critical. We also have the name of the technology affected, the name of the vendor of the technology, and the EPSS rating, which indicates the likelihood of this vulnerability being exploited in the wild.

On the second area, we present the exploits found in Exploit DB if any are found (Figure 3). This tab returns to the user, the ID of the exploit on the catalog, the author of the exploit, the description of the exploit, the date on which the exploit was published, the addresses to the exploit on the catalog and the URLs to download the exploit.

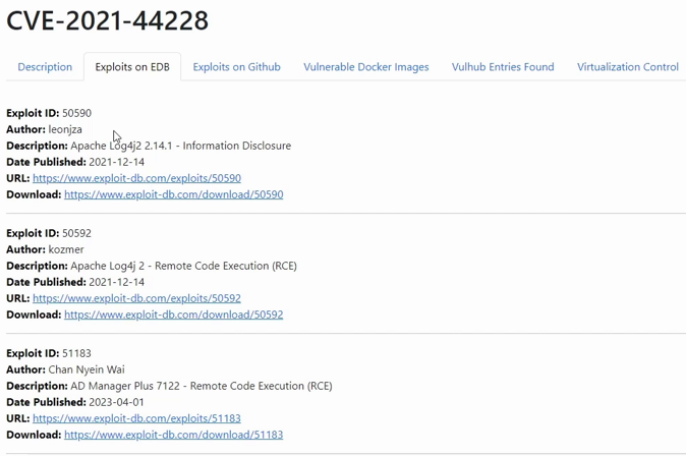


Figure 3. Exploits found on Exploit-DB

In the next area, we have the potential PoC exploits found on Github, if any are found. The repository's number of stars orders these repositories on the results. The information retrieved includes the name of the repository, the author of the repository, the description of the repository, the date on which the repository was created, the URL for the repository, and the number of stars the repository has.

The fourth area contains the potential purposely built vulnerable docker images from Docker Hub[9] if any are found. Here, users can find the vulnerable docker images sorted by the number of pull counts. On this page, we can find information for the docker images such as the name of the image, the name of the author of the docker image, the description of the docker image, the URL to the image on docker hub and a button to deploy this docker image on the VM.

The fifth area contains an entry to the Vulhub repository[10] if it exists. Here users can get the URLs to the entry on the Github of Vulhub and a button to deploy the vulnerable docker image on the local VM.

On the last area, we have the virtualization control interface, with buttons that control a Kali Linux VM. Users can start, and stop the VMs, have access to VNC and SSH connection in the web-browser to the VMs. In this area, we also have a VNC connection to the Kali VM. Besides that, it is also possible to control the Ubuntu Server that hosts the Docker Containers, with similar actions. We also got the status of both

[8]https://github.com/vulhub/vulhub

[9]https://hub.docker.com/

[10]https://github.com/vulhub/vulhub/

VMs, if they are running or not running, and for how long. Besides that, on this page, we also have control for the docker containers. We can get a file with the capture of packets from the docker container, and we are also able to stop the docker containers and have information such as if docker containers are running or not, the IP of the machine with the docker containers, and the ports opened by docker.

B. Virtualization

To perform the needed virtualization support for the, as a base deployment we consider three VMs. The first one is a Kali Linux VM, from which the attacks will be launched towards vulnerable applications. The second one is an Ubuntu server, that will serve as the machine that will host the vulnerable docker containers. It will also contain a docker container that is gathering network logs, from the vulnerable docker container.

The third one is another Ubuntu Server, that will serve as the entry point to the network, this Ubuntu server will host a WebSSH instance and a NoVNC instance, from these services the user will then be able to interact with the attacker VM and the victim VM.

IV. Results

In order to test the application, we executed the different workflows, in which we attempted to mimic the actions that would be performed by a security analyst.

A. Application Workflow with Popular Vulnerability

The first workflow we did was starting the app, searching for a popular CVE of a popular technology, start a vulnerable docker image from Docker Hub to this CVE, and try a PoC on Exploit-DB against it, and then get a network capture of the traffic for the docker container and also get some logs from the docker container.

To do that, we started the application, and opened our browser and visited the application home page. On that page, we searched for the popular technology we wanted to evaluate, in tihs case `log4j`. We obtained the CVEs to the searched technology on that same page. We then selected the one we wanted, CVE-2021-44228, which was, and still is, a popular CVE. This search is depicted in Figure 4. Any other CVE could be used as we index all existing CVEs.

Being on the page for the CVE, we had the opportunity to take notice of the pieces of information about that CVE. We navigated to the tab “Exploits on EDB”, there, we looked for the exploits and opened a new tab in our browser, one of those URLs for an exploit. Then we navigated to the tab Vulnerable Docker Images, which can be seen in Figure 5, and selected the button to deploy one of those images using a docker container. After waiting a brief moment, it deployed the docker image.

Next, in this workflow, we navigated to the tab “Virtualization Control”, were we selected the button “Show Attacker SSH” to open an SSH session on our browser inside the attacker’s machine. Having that session opened, we obtained the exploit previously identified. We then followed the instructions to execute the exploit. We can view a SSH session to inside attacker VM.

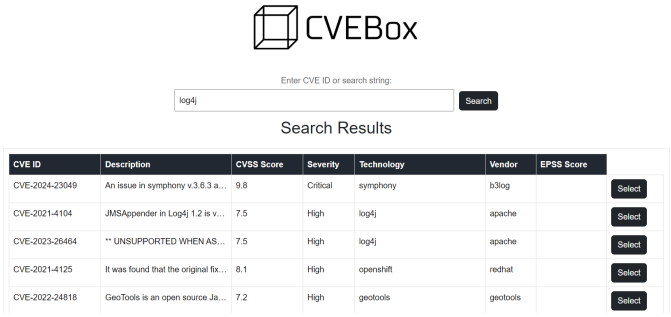


Figure 4. Searching log4j on the application

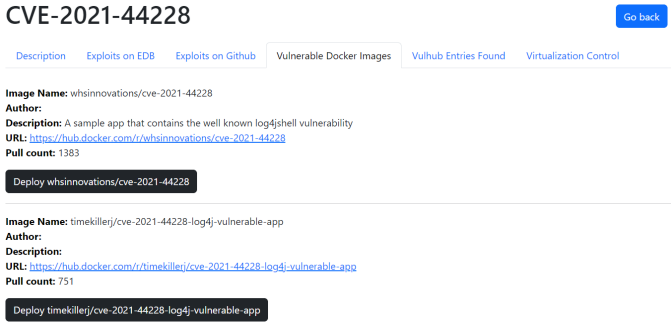


Figure 5. Tab with vulnerable docker images

Then, we performed an analysis of the exploit executed. To do that, we went back to the tab “Virtualization Control” on our application, and in the “Docker Container Control” section, we clicked the button “Get Network Logs of Docker”. This immediately downloaded a packet capture of the logs for the docker container, which we were able to open on our machine using Wireshark. The result is depicted in Figure 6

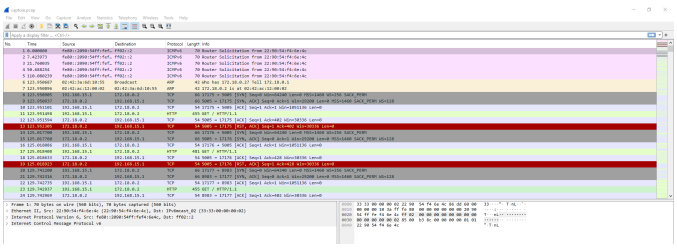


Figure 6. Packet capture obtained from the docker container

After that we also wanted the logs for the application (in this case it’s Tomcat) of the exploited docker container. To do that, we went back to the tab “Virtualization Control” of our application, and on the section “Docker Host Server Control”, we selected the button “Show Docker Host SSH”, which opened an SSH session inside the Docker host server. There we can obtain the docker containers running, and execute commands

inside it using the standard docker interface. Having that shell opened, we can use cat or other command to inspect the logs we wanted to look into. In addition the analyst, can inspect the logs of the container it self, or from other solutions (e.g. syslog) if the deployment is built that way.

In this workflow, we had the opportunity to have an SSH session inside the attacker machine to explore the CVE, using the availablePoC.

We also had the opportunity to take a look at a packet capture obtained from using the web application, and access the victim machine via SSH to obtain logs.

B. Application Workflow using Vulhub entry

The second workflow we performed was similar to the one before, but this time, we wanted to explore a CVE that had an entry in the Vulhub Github Repository. Since this repository is a popular one, we can expect that when an entry is added to that repository, many people will learn how CVEs are explored, including malicious actors.

This application can search entries and deploy vulnerable containers from that GitHub repository. To do that, similar to before, we searched for a CVE that we know exists in that repository (CVE-2023-29300). Being on the page for the CVE, we selected the tab "Vulhub Entries Found". We opened the URL entry on a new tab in our browser and we used the interface to deploy the vulnerable docker image associated with the entry.

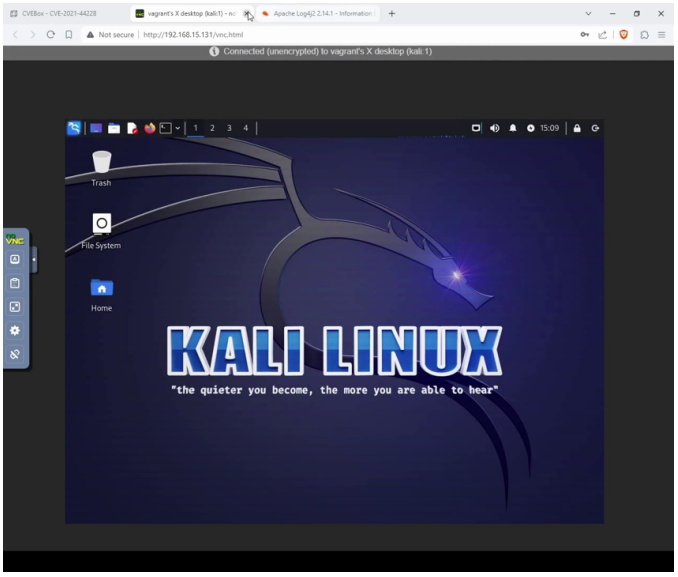


Figure 7. VNC Connection to inside Kali Linux

After waiting for the docker container to be deployed, we selected the tab "Virtualization Control" on our application, and this time, we decided to explore the PoC exploit using the VNC connection to our attacker machine. We opened a terminal on the VNC connection to the attacker machine (Figure 7). We obtained the exploit from the Vulhub Github repository using the provided URL, and then we followed the steps on the repository to explore the vulnerability.

In the third workflow, we wanted to get information and all the possible PoC and vulnerable docker containers that could exist or not for a new CVE. To do that, we started the app as in the previous two workflows, scrolled to the section Latest Updated Critical CVEs, and selected some of those CVEs. Having navigated on the tabs inside the application ("Exploits on EDB", "Exploits on Github", "Vulnerable Docker Images", and "Vulhub Entries Found"), we noticed that many of them had no entries on those areas. That is because they are new and there is little knowledge about them. We obtained one or two possible PoC and some information about this new CVE.

With these tests we functionally evaluated the solution, demonstrating that it provides the expected value, allowing analysis to rapidly deploy, and evaluate existing CVEs, even if new.

C. Performance Evaluation

We also executed a performance evaluation of the time taken by the application to perform numerous actions. This noted times are registered in Table I.

Action	Time (sec)
Starting the Application	410
Searching CVEs by technology name	1
Loading CVE page never loaded before	45
Loading CVE page loaded before	12
Deploy a docker image from Docker Hub	10-15
Deploy a docker image from Vulhub	10-15
Stop docker containers	7
Stop attacker machine	8
Start attacker machine	165
Stop Docker Host machine	42
Start Docker Host machine	82
Get Network Logs of Docker	3
Stop the application	50

Table I
TIME TAKEN FOR ACTIONS ON THE APPLICATION

As show, the execution time is measured in seconds for all tasks, and within reasonable limits for the use case. The times are highly dependent on the hardware used, and the virtualization platform in use to instantiate the Virtual Machines. Without the tool we present, a member of the Incident Response team would require tens of minutes to validate a CVE on existing software. We now allow a dramatic reduction, in the effort and time required, which is placed at little more than 5 minutes. It should be noticed that the environment automatically captures software interactions, allowing to rapidly detect how an exploit is conducted, and to validate mitigation strategies.

V. CONCLUSION

We have observed that cybersecurity is an expanding field with the need for more and more professionals to help address cyber threats. We have also observed the transition from a classic learning approach in the field to a more hands-on and gamified learning approach. We also explored the appearance of standards such as CVEs, and their importance to the field.

In this work, we proposed a tool to solve a problem experienced by professionals, that is the difficulty of learning about CVEs, by creating a platform that attempts to teach about the CVEs in a more modern and hands-on manner by using VMs to explore the CVE directly in a web application, compared to the classic way that is reading a text and having to look around the web for possible PoCs of the CVE and having to create a manual lab to test the CVE.

The authors defend that platforms such as the one created in this paper will be created and widely used based on the need for more professionals and the sheer amount of CVE appearing each day. Faster and better learning approaches, such as hands-on and gamified learning, are the way forward to learning in cybersecurity and especially to learning about CVEs.

Even though the platform developed here is just a simple concept of what it could be, we hope we have contributed with at least an idea of what could be developed in the future and its impact.

VI. Acknowledgments

This work was supported by FCT - Fundação para a Ciência e Tecnologia, I.P. by project reference UIDB/50008: Instituto de Telecomunicações.

References

[1] Thomas M. Chen and Saeed Abu-Nimeh. Lessons from stuxnet. *Computer*, 44:91–93, 2011. doi:10.1109/MC.2011.115.

[2] Graeme R Newman, Megan M McNally, and other. Identity theft literature review. Literature Review 210459, National Institute of Justice United States Department of Justice, 2005. URL https://nij.ojp.gov/library/publications/identity-theft-research-review.

[3] Peter Mell and Tim Grance. Use of the common vulnerabilities and exposures (cve) vulnerability naming scheme. *NIST Special Publication*, 800:51, 2002.

[4] Soufian El Yadmani, Robin The, and Olga Gadyatskaya. How security professionals are being attacked: A study of malicious cve proof of concept exploits in github. *arXiv preprint arXiv:2210.08374*, 2022. doi:10.48550/arXiv.2210.08374.

[5] Sujit Subhash and Elizabeth A Cudney. Gamified learning in higher education: A systematic review of the literature. *Computers in human behavior*, 87:192–206, 2018.

[6] K Boopathi, S Sreejith, and A Bithin. Learning cyber security through gamification. *Indian Journal of Science and Technology*, 8(7):642–649, 2015. doi:10.17485/ijst/2015/v8i7/67760.

[7] Ehsan Aghaei and Ehab Al-Shaer. Cve-driven attack technique prediction with semantic information extraction and a domain-specific language model, 2023. URL https://arxiv.org/abs/2309.02785.

[8] Heedong Yang, Seungsoo Park, Kangbin Yim, and Manhee Lee. Better not to use vulnerability's reference for exploitability prediction. *Applied Sciences*, 10(7), 2020. ISSN 2076-3417. doi:10.3390/app10072555.

[9] Jan Vykopal, Pavel Celeda, Pavel Seda, Valdemar Svábenský, and Daniel Tovarnák. Scalable learning environments for teaching cybersecurity hands-on. *CoRR*, abs/2110.10004, 2021.

AI-aided compost by digital twins: A revolutionary symbiosis or an overengineered dream?

Jorge Cancho Casado
COMPUTAEX Foundation
Cáceres, Spain
jorge.cancho@computaex.es

Marcos Jesús Sequera Fernández
Grupo de Ingeniería de Medios (GIM)
Cáceres, Spain
marcosjesus@unex.es

Fernando Broncano Morgado
Grupo de Ingeniería de Medios (GIM)
Cáceres, Spain
fbroncano@unex.es

Antonio Gordillo Guerrero
Statistical Physics In Extremadura
Cáceres, Spain
anto@unex.es

Pablo García Rodríguez
Grupo de Ingeniería de Medios (GIM)
Cáceres, Spain
pablogr@unex.es

***Abstract*—Despite recent advancements, Spain and Europe still face major challenges in meeting the European Union's recycling targets for 2025 and beyond. To address this, the paper presents a digital twin system that aims to replicate the physical composting process in real time. This system allows users to visualize and predict the compost state by adjusting environmental parameters such as light, humidity, and temperature. Leveraging Internet of Things (IoT), cybersecurity and artificial intelligence (AI), it gathers sensor data from the physical composter to simulate and optimize the composting process. The digital twin enhances process efficiency, enables predictive analysis, supports policy-making, and contributes to circular economy goals. The paper discusses the system's design, implementation, and potential for adoption in smart waste management.**

***Index Terms*—Digital Twins, Composting, IoT, Cybersecurity, Artificial Intelligence, Organic waste**

I. Introduction

The management of organic waste has become a critical challenge in the pursuit of sustainable urban development and the transition towards a circular economy. Despite significant advancements in waste management practices, both Spain and Europe at large are grappling with the complexities of meeting the ambitious recycling targets set by the European Union for 2025 and beyond. The current state of organic waste recycling in Europe paints a concerning picture: a mere 17% of municipal solid waste is organically recycled through composting and anaerobic digestion. More alarmingly, approximately 74% of kitchen waste still finds its way to landfills or incinerators, representing a significant loss of potential resources and contributing to greenhouse gas emissions [1].

In Spain, recent regulations have mandated that by 2025, all businesses must ensure the separation of waste at source, including categories such as organic waste, plastic, paper and cardboard, glass, and metals. This legislative push aims to align with the EU's objectives of recycling at least 55% of municipal waste by 2025, with subsequent targets of 60% by 2030 and 65% by 2035. However, the path to achieving these goals is fraught with challenges, including inadequate infrastructure, complex local recycling systems, and the need for increased public awareness and participation [2].

The main contribution of this project is the ability to simulate behavior and display the entire composting process.

II. Digital Twins

To address these issues, this paper presents the development of a digital twin for a composting system. This technology creates a virtual replica of the physical composting process, allowing users to visualize and predict compost state in real-time by manipulating key environmental parameters such as light, humidity, and temperature (Figure 1). By leveraging the Internet of Things (IoT), cybersecurity and artificial intelligence (AI), the system processes data from sensors embedded in the physical composter to create an accurate and dynamic virtual representation.

Digital twin technology offers important advantages for composting by enabling real-time monitoring and adjustment of critical parameters, which improves both the efficiency and quality of compost production. It supports predictive analysis and process control through scenario simulation, helping reduce compost maturation time and enhance the final product. Additionally, its visual and interactive features promote public engagement and education on sustainable waste management. Finally, it provides valuable data to support policy implementation and planning, especially in the context of meeting EU recycling targets.

III. Proposed Methodology

The mathematical modeling underpinning the digital twin incorporates complex equations describing the biological, chemical, and physical processes occurring during composting. This includes mass and energy balances, kinetic models of microbial activity, temperature evolution, moisture dynamics, and substrate degradation. The resulting system of differential equations is solved using advanced numerical methods, allowing for accurate simulation and

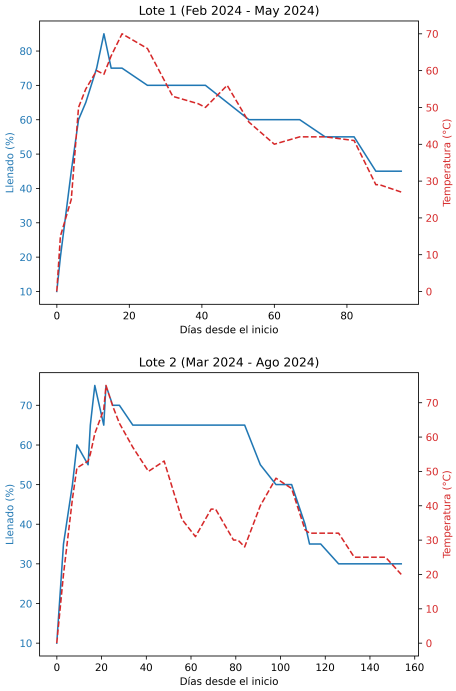


Fig. 1. Representation of temperature and filling level trends in two composters throughout the composting period.

prediction of compost behavior under various conditions [3]. Several dynamic and structured models have also demonstrated accurate simulations validated through experimental composting setups [4], [5].

Implementing this digital twin system demands a robust technological infrastructure, including sensor networks, cloud computing, cybersecurity and advanced data analytics. Sensors embedded in composting units collect real-time data (temperature, humidity, oxygen, pH) and send it to a cloud-based platform. This platform processes and integrates the data using machine learning algorithms to refine the digital model and improve predictive accuracy. Recent works confirm the relevance of machine learning models in predicting compost maturity and optimizing composting strategies [6], [7]. The combination of real-time monitoring and AI-driven insights enables dynamic process control, ensuring optimal composting conditions at all times. Knowledge-based control of composting parameters has been shown to significantly improve process performance and efficiency [8].

A digital model of a composter was developed using Unity [9] (Figure 2) to simulate the conditions of a real composting system. This model represents the progression of the composting process over time, displaying key parameters such as temperature and overall mixture conditions. It allows users to visualize how the compost evolves and responds to different environmental factors. When suboptimal conditions are detected, the system suggests appropriate corrective actions to maintain effective composting.

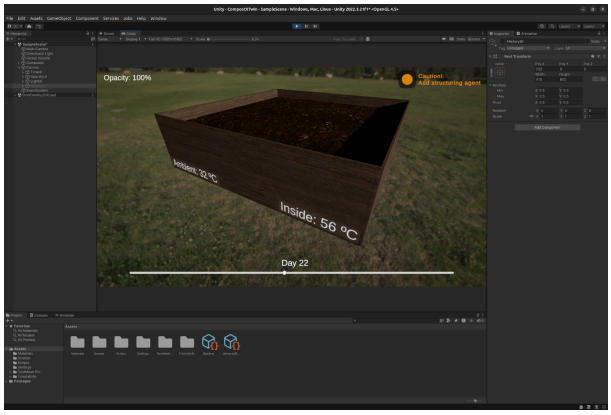


Fig. 2. Virtual model of the composter developed in Unity.

IV. Future Goals

Upcoming studies should focus on implementing physical behavior on this model, as well as sensing the devices through real-time interaction. Next steps should prioritize cost-effective sensor solutions, data standardization, and the development of scalable business models to facilitate widespread adoption.

Acknowledgment

This initiative is carried out within the framework of the funds of the Recovery, Transformation and Resilience Plan, financed by the European Union (Next Generation) - National Cybersecurity Institute (INCIBE) in the project C108/23 “Detection of Identity Document Forgery Using Computer Vision and Artificial Intelligence Techniques”.

Thank you also to Pontevedra Provincial Council (Deputación de Pontevedra, Spain) for granting access to their compost data.

References

[1] “Ecn data report 2022 compost and digestate for a circular bioeconomy.” [Online]. Available: https://www.compostnetwork.info/wordpress/wp-content/uploads/ECN-rapport-2022.pdf

[2] “Bio-waste in europe-turning challenges into opportunities,” 2020. [Online]. Available: https://www.eea.europa.eu/en/analysis/publications/bio-waste-in-europe

[3] E. Walling, A. Trémier, and C. Vaneeckhaute, “A review of mathematical models for composting,” *Waste Management*, vol. 113, pp. 379–394, 7 2020.

[4] I. Petric and V. Selimbašić, “Development and validation of mathematical model for aerobic composting process,” *Chemical Engineering Journal*, vol. 139, pp. 304–317, 6 2008.

[5] A. P. Gomes and A. F. Pereira, “Mathematical modelling of a composting process, and validation with experimental data,” *Waste Management and Research*, vol. 26, pp. 276–287, 6 2008.

[6] S. Shi, Z. Guo, J. Bao, X. Jia, X. Fang, H. Tang, H. Zhang, Y. Sun, and X. Xu, “Machine learning-based prediction of compost maturity and identification of key parameters during manure composting,” *Bioresource Technology*, vol. 419, p. 132024, 3 2025.

[7] S. Ding and D. Wu, “Comprehensive analysis of compost maturity differences across stages and materials with statistical models,” *Waste Management*, vol. 193, pp. 250–260, 2 2025.

[8] “Knowledge is power in compost process control biocycle.” [Online]. Available: https://www.biocycle.net/knowledge-is-power-in-compost-process-control/

[9] “Unity real-time development platform — 3d, 2d, vr & ar engine.” [Online]. Available: https://unity.com/

Imbalance-Aware Intrusion Detection with a Two-Stage Binary Classification System

Óscar Mogollón-Gutiérrez
Universidad de Extremadura
oscarmg@unex.es

Marcos J. Sequera Fernández
Alberto López-Trigo
Universidad de Extremadura
{marcosjesus,albertolt}@unex.es

Mohammadhossein Homaei
Universidad de Extremadura
mhomaein@alumnos.unex.es

José Carlos Sancho Núñez
Universidad de Extremadura
jcsanchon@unex.es

***Abstract*—Cyberattacks are growing in both frequency and complexity, prompting the need for effective cybersecurity measures. Intrusion detection systems (IDS), especially when combined with artificial intelligence (AI), play a vital role in identifying malicious network activity. This paper proposes a two-step IDS that uses binary classifiers—one per attack type—to first detect, then categorize cyberattacks. The method addresses class imbalance via oversampling and was evaluated on two intrusion detection datasets. It outperformed classical and state-of-the-art models, achieving F1-scores of 0.7213 (NSL-KDD) and 0.9793 (ToN-IoT).**

***Index Terms*—cyberattack detection, imbalance learning, IDS**

I. Introduction

Recent research on intrusion detection systems (IDS) has explored a range of methods including traditional single classifiers, ensemble learning, and deep learning approaches. Multi-model learning and two-stage classification proposal have shown promise for enhancing detection accuracy. Notably, the One-vs-Rest (OvR) strategy has been used for imbalanced multiclass problems, and techniques like SMOTE have been employed to address class imbalance [1]. However, most existing work does not integrate OvR with a comprehensive binary classification system that simultaneously enhances detection and classification under severe imbalance. Our approach fills this gap by introducing a two-stage binary model architecture optimized with SMOTE, evaluated across multiple types of attacks.

II. Related works

With the proliferation of connected systems and sophisticated cyber threats, intrusion detection systems (IDS) have evolved significantly. Traditional approaches include binary and multiclass classification using models such as decision trees and support vector machines. For instance, Khammassi [2] applied a genetic algorithm for feature selection with decision trees, while recent work has explored deep learning techniques like CNNs [3].

Ensemble and hybrid models have been proposed to improve detection accuracy and generalizability [4]. Despite their effectiveness, these models often overlook class imbalance, a persistent issue in IDS where malicious traffic forms the minority. To address this, techniques such as SMOTE, undersampling, and cost-sensitive learning have been employed [5].

However, many studies remain limited to single classifiers or binary detection. Few adopt staged or multi-model approaches. This work addresses that gap by applying SMOTE-based One-vs-Rest classifiers for each class, evaluated using imbalance-aware metrics across multiple datasets.

III. Method

This section presents the proposed two-stage binary classification system for intelligent network traffic analysis. The methodology consists of three main phases: data preprocessing, binary model generation, and model integration.

A. An overview of proposed intrusion detection system

After preprocessing the raw traffic data for learning algorithms, we construct binary classifiers—each trained to recognize a specific attack type versus all others. These models are then integrated into a multi-model architecture capable of detecting and classifying cyberattacks.

For each attack category, a balanced binary dataset is created using oversampling (e.g., SMOTE) to address class imbalance. Each dataset includes positive samples (target category) and negative samples (a balanced mix of all other categories). This ensures that the classifiers are trained to detect minority class patterns effectively.

Each binary classifier is optimized using grid search and 5-fold cross-validation across four algorithms: KNN, SVM, Decision Trees, and MLP. Macro-F1 score is used for evaluation, as it equally emphasizes both classes in the binary setting.

B. Two-Stage system design

Figure 1 summarizes the architecture. In the first stage, a binary classifier distinguishes between benign and malicious traffic. If traffic is flagged as suspicious, it proceeds to the second stage, where each class-specific model outputs its confidence score. The class with the highest score is selected as the predicted attack type.

This OvR-based proposal enables fine-grained classification while preserving detection sensitivity. The design also incorporates oversampling and hyperparameter tuning to enhance robustness.

C. Experimental setup

To validate the proposed framework, we used two widely adopted intrusion detection datasets: NSL-KDD,and ToN-IoT. These datasets reflect diverse attack types and traffic patterns, while also exhibiting significant class imbalance.

Four classifiers were evaluated: K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Trees (DT), and Multilayer Perceptron (MLP). Hyperparameter tuning was performed using grid search with 5-fold cross-validation, optimizing for the Macro-F1 score to ensure balanced performance across classes.

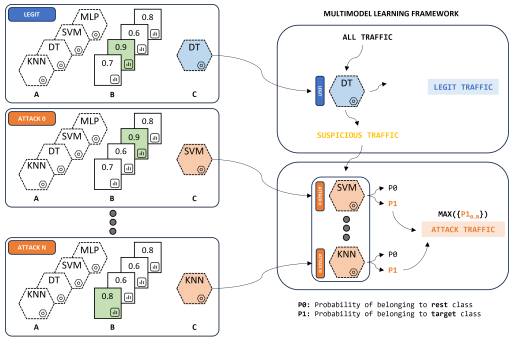


Fig. 1. Two-stage binary classification System

IV. Experimental results

Performance comparison of the proposed model is presented in this section using four widely used cybersecurity datasets: NSL-KDD TON-IoT. The proposal is evaluated using well-known classification metrics and compared with other research proposed in the scientific literature.

A. Experimentation and discussion on NSL-KDD

First, the proposed model was evaluated using the NSL-KDD dataset. To accomplish this, our approach combines the best binary models for each of the five types of traffic collected: DoS, Probe, R2L, U2R, and Normal.

Our proposal achieves an F1 of 0.7213, higher than the best F1 score of 0.7126 obtained by another scientific work in intrusion detection using the NSL-KDD dataset. This suggests that our model can achieve better performance in terms of the balance between precision and recall compared to the other works.

B. Experimentation and discussion on ToN-IoT

The final set of experiments involved testing the proposed model on the ToN-IoT set. It was necessary for the proposal to distinguish different categories of traffic in a network of IoT devices.

A comparison is made between the results obtained in this research and previous contributions to the detection of intrusions on ToN-IoT. In terms of evaluation metrics, the proposal improved both works. Additionally, it performed better than the individual classifiers. In the same manner as with the category performance, the DT performs significantly better than the SVM, MLP, and KNN.

TABLE I
Comparison of performance against state-of-art works

Work	Dataset	Accuracy	Precision	Recall	F1
[6]	NSL-KDD	**0.7660**	-	-	-
[3]	NSL-KDD	0.7608	0.7360	0.7608	0.7126
Proposed	NSL-KDD	0.7620	0.7329	0.7620	**0.7213**
[7]	TON-IoT	0.9380	-	0.9229	0.9400
[8]	TON-IoT	-	0.9730	0.9610	0.9660
Proposed	TON-IoT	**0.9788**	**0.9807**	**0.9788**	**0.9793**

V. Conclusion

This paper presents a binary model learning framework for cyberattack detection and classification in networked environments, addressing class imbalance through minority oversampling. Evaluated on two benchmark datasets (NSL-KDD and ToN-IoT), the method consistently outperforms existing approaches in F1-score, demonstrating robust classification under imbalance. Future work includes exploring alternative oversampling techniques and extending the system to anomaly detection in heterogeneous domains.

VI. Acknowledgements

This initiative is carried out within the framework of the funds of the Recovery, Transformation and Resilience Plan, financed by the European Union (Next Generation) - National Cybersecurity Institute (INCIBE) in the project C108/23 “Detection of Identity Document Forgery Using Computer Vision and Artificial Intelligence Techniques”

References

[1] O. Mogollón-Gutiérrez, J. C. Sancho Núñez, M. Ávila, and A. Caro, “A detailed study of resampling algorithms for cyberattack classification in engineering applications,” *PeerJ Computer Science*, vol. 10, p. e1975, Apr. 2024.

[2] C. Khammassi and S. Krichen, “A GA-LR wrapper approach for feature selection in network intrusion detection,” *Computers and Security*, vol. 70, pp. 255–277, 2017. [Online]. Available: https://doi.org/10.1016/j.cose.2017.06.005

[3] W. Wang, X. Du, D. Shan, R. Qin, and N. Wang, “Cloud intrusion detection method based on stacked contractive auto-encoder and support vector machine,” *IEEE Transactions on Cloud Computing*, vol. 10, no. 3, pp. 1634–1646, Jul. 2022. [Online]. Available: https://doi.org/10.1109/tcc.2020.3001017

[4] T. Al-Shehari and R. A. Alsowail, “Random resampling algorithms for addressing the imbalanced dataset classes in insider threat detection,” *International Journal of Information Security*, Dec. 2022. [Online]. Available: https://doi.org/10.1007/s10207-022-00651-1

[5] M. Rani and Gagandeep, “Effective network intrusion detection by addressing class imbalance with deep neural networks multimedia tools and applications,” *Multimedia Tools and Applications*, vol. 81, no. 6, pp. 8499–8518, Feb. 2022. [Online]. Available: https://doi.org/10.1007/s11042-021-11747-6

[6] G. Andresini, A. Appice, and D. Malerba, “Autoencoder-based deep metric learning for network intrusion detection,” *Information Sciences*, vol. 569, pp. 706–727, Aug. 2021. [Online]. Available: https://doi.org/10.1016/j.ins.2021.05.016

[7] M. Sarhan, S. Layeghy, and M. Portmann, “Evaluating standard feature sets towards increased generalisability and explainability of ML-based network intrusion detection,” *Big Data Research*, vol. 30, p. 100359, Nov. 2022. [Online]. Available: https://doi.org/10.1016/j.bdr.2022.100359

[8] A. Telikani, J. Shen, J. Yang, and P. Wang, “Industrial IoT intrusion detection via evolutionary cost-sensitive learning and fog computing,” *IEEE Internet of Things Journal*, vol. 9, no. 22, pp. 23 260–23 271, Nov. 2022. [Online]. Available: https://doi.org/10.1109/jiot.2022.3188224

The Role of Physical Personal Identification in Documentary Cybersecurity

Fernando Broncano Morgado
Grupo de Ingeniería de Medios
Universidad de Extremadura
fbroncano@unex.es

Victoria Amores Chaparro
Grupo de Ingeniería de Medios
Universidad de Extremadura
vicamores@unex.es

José Carlos Sancho Núñez
Grupo de Ingeniería de Medios
Universidad de Extremadura
jcsanchon@unex.es

***Abstract*—The emergence of artificial intelligence –AI– has significantly impacted personal identification systems. This work proposes an integrated pipeline based on AI and computer vision to detect forgeries in digital images of identity documents. The system evaluates security elements, facial validity, and personal data redundancy using advanced techniques. This pipeline aims to ensure document integrity and mitigate presentation attacks, aligning with the documentary cybersecurity of information systems.**

***Index Terms*—Documentary cybersecurity, identification documents, forgery detection, presentation attacks, OCR, ORB, ViT**

I. Introduction

The emergence of Artificial Intelligence in society has brought about significant changes that have influenced people's behavior. Similarly, the COVID-19 pandemic has also conditioned various tasks that people used to perform under previous habits. One of the habits most affected is personal identification.

Traditionally, personal identification has been carried out through physical mechanisms, such as the earliest forms of safe-conducts, which facilitated the free movement of individuals, and various documents designed to verify that the bearer is indeed who they claim to be. As society has advanced, the use of identity documents, visas, and passports has become standardized, allowing citizens to move freely across different economic and social spaces.

This personal identification also differs in its application scope. Countries with legislation based on the Anglo-Saxon law have recently required the introduction of documents to identify their citizens within their own borders, such as the United States or the United Kingdom [1], [2]. Other countries, including Spain or Portugal, or the majority of European countries, have enabled mechanisms to allow the identification of individuals for a couple of centuries.

In this regard, and with the rise of aviation, globalization, and the free movement of people, important precedents have been set for the standardization of international travel documents. In this context, the International Civil Aviation Organization –ICAO– has established standards for the creation of identification documents at an international level. These standards ensure that all documents meet uniform conditions, including basic data to be collected, their structure, and features to facilitate recognition. These recommendations are outlined in the ICAO-9303 standard [3]. This standard refers to these identifiers as travel documents, which are divided into three categories: (1) Type 1 travel documents *–such as Spanish National Identity Cards–*; (2) Type 2 travel documents *–related to visa issuance–*; and Type 3 travel documents *–reserved for passport structure–*.

Travel documents incorporate various mechanisms for the easy verification of data. These documents are divided into two zones: the visual inspection zone –VIZ–; and the machine-readable zone –MRZ–. The visual inspection zone must allow data to be verified with a simple glance by a human, while the machine readable zone emerged with advancements in optical character recognition –OCR– and computer vision. Early advancements in this field enabled the mechanized reading of information through a simple scan of these lines, which summarize the most critical data. Likewise, identity documents have gradually been digitized and stored as images.

The ease of use of photo editing software has increasingly driven the need to verify and cross-check the data on identity documents. Initially, these modifications were crude, with the original areas clearly distinguishable from the altered ones. However, these alterations have become increasingly sophisticated, resulting in documents that appear visually correct.

With the introduction of Artificial Intelligence, these alterations have increasingly shifted toward modifying the face of the person identified in the document. This allows the face in the document to be physically associated with another individual, effectively attributing the document's data to a person with a different physical appearance. These issues can also be applied to other types of documents, such as driver's licenses, enabling the identity of someone without a driving permit to be usurped by another person who does possess one.

All these attacks on information systems fall under the category of presentation attacks. In these attacks, the appearance of the document representing the physical identity such as a digital image is modified in order to carry out intrusions into systems. Examples of systems that could be targeted by these intrusions include online banking entities, rental contract signings, or airplane boarding. These systems must establish strong authentication mechanisms and verification processes for the documents uploaded.

In the literature, various techniques employed to carry out presentation attacks on identity documents from different

countries can be found [4]–[9].

This work proposes the introduction of an integrated processing pipeline for the detection of counterfeit Identity Documents, in order to ensure the immutability and integrity of the documents.

II. Proposed Processing Pipeline

In this section, a brief introduction to the security features implemented in the construction of more secure Identity Documents is presented. Similarly, the pipeline for the detection of counterfeits and the features to be applied to the counterfeiting of these documents are explained.

A. Security features in identity documents

To enhance security and make the creation of fraudulent identity documents more difficult, the European Union has compiled all applicable security measures in a glossary of terms known as PRADO [10]. Similarly, in the same vein, European Directive 2019/1157 [11] urges EU member states to improve the security of their documents by standardizing certain terms, introducing national emblems, and incorporating some of PRADO's security features.

Some of the security features incorporated in identity documents include ultraviolet inks, optically variable inks, transparent windows, embossed inscriptions, or a hologram of national symbols. To detect forgeries, it will be necessary to verify that the images contain these elements represented.

B. Forgery detection pipeline

A forgery detection pipeline must examine both the information and the format of the identity document's representation in a digital image. As a first step, the various security features present in an identity document, such as the Kinegram, must be detected within the image. To achieve this, three alternatives are proposed: (1) performing a hyperspectral analysis to study the behavior of different inks; (2) building a model based on a Visual Transformer [12] trained with authentic documents to classify their authenticity; (3) verifying its genuineness by extracting features *–using ORB [13]–* and comparing them against a genuine identity document.

As a second step, the evaluation of the document's face is proposed to verify its authenticity, following a methodology such as the one presented in [14]. Finally, an analysis of data redundancy is performed, ensuring that the information in the visual inspection zone and the machine-readable zone is consistent. Using OCR, text strings can be extracted from the identity document, enabling the digitization of the data.

This three-step pipeline is based on the analysis of the image, the person's face, and the examination of the personal data. With this approach, it is possible to verify authenticity and process the digitization of the document's data.

III. Conclusions

The methodology proposed in this work aims to achieve a system based on artificial intelligence and computer vision qualified to detect fraudulent identity documents in their digital image representation. Specifically, this methodology must compare the validity of the data, its homogeneity, and the documentary integrity in the digital image representation. In this way, it seeks to avoid presentation attacks in information systems that require personal identification using physical documents.

Acknowledgements

This initiative is carried out within the framework of the funds of the Recovery, Transformation and Resilience Plan, financed by the European Union –Next Generation– and National Cybersecurity Institute –INCIBE– in the Project C108/23 "Detección de Falsificación de Documentos de Identidad mediante Técnicas de Visión por Computador e Inteligencia Artificial".

References

[1] D. J. Haas and B. Zimmer, *Personal Identification: Modern Development and Security Implications, Second Edition*. Personal Identification: Modern Development and Security Implications, Second Edition, 2024.

[2] I. Liersch, *ID cards and passports*. Smart Cards, Tokens, Security and Applications: Second Edition, 2017.

[3] *Doc 9303 Documentos de viaje de lectura mecánica*, Organización de la Aviación Civil Internacional, 2021. [Online]. Available: https://www.icao.int/publications/pages/publication.aspx?docnum=9303

[4] D. Benalcazar, J. E. Tapia, S. Gonzalez, and C. Busch, "Synthetic id card image generation for improving presentation attack detection," *IEEE Transactions on Information Forensics and Security*, 2023.

[5] D. Bothra, S. Dixit, D. P. Mohanty, M. Haseeb, S. Tiwari, and A. Chaulwar, "Synthetic data generation pipeline for private id cards detection," in *2023 IEEE Women in Technology Conference (WINTECHCON)*, 2023.

[6] V. Amores Chaparro, "Técnicas avanzadas para la generación y detección de Documentos de Identidad sintéticos," Master's thesis, Escuela Politécnica de Cáceres, Universidad de Extremadura, 2022.

[7] J. E. Tapia, N. Damer, C. Busch, J. M. Espin, J. Barrachina, A. S. Rocamora, K. Ocvirk, L. Alessio, B. Batagelj, S. Patwardhan, R. Ramachandra, R. Mudgalgundurao, K. Raja, D. Schulz, and C. Aravena, "First competition on presentation attack detection on id card," in *2024 IEEE International Joint Conference on Biometrics (IJCB)*, 2024.

[8] V. Amores Chaparro, F. Broncano Morgado, Hernández Martín, Mogollón Gutiérrez, J. C. Sancho Núñez, and S. Pais, "Generación de documentos nacionales de identidad sintéticos mediante el uso de perfiles biográficos," in *IX Jornadas Nacionales de Investigación En Ciberseguridad*, 2024.

[9] A. Sanchez, J. M. Espín, and J. E. Tapia, "Few-shot learning: Expanding id cards presentation attack detection to unknown id countries," in *2024 IEEE International Joint Conference on Biometrics*, 2024.

[10] *Glorario de medidas de seguridad de documentos de identidad*, Secretaría General del Consejo de la Unión Europea, 2022. [Online]. Available: https://www.consilium.europa.eu/prado/es/prado-glossary/prado-glossary.pdf

[11] *Reglamento UE 2019/1157*, Boletín Oficial del Estado, 2019. [Online]. Available: https://www.boe.es/buscar/doc.php?id=DOUE-L-2019-81170

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[13] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International Conference on Computer Vision*, 2011, pp. 2564–2571.

[14] F. Broncano Morgado, V. Amores Chaparro, Mogollón Gutiérrez, A. López Trigo, and J. C. Sancho Núñez, "Evaluación de rostros faciales para la expedición de documentos de identificación," in *XVIII Reunión Española sobre Criptología y Seguridad de la Información*, 2025.

Snakey: A Blue Team Keylogger for Insider Threat Detection

Martim L. S. Rebelo
Department of Computer Science
Universidade da Beira Interior
Covilhã, Portugal
martim.l.rebelo@ubi.pt

João B. F. Sequeiros
Department of Computer Science
Instituto de Telecomunicações, Universidade da Beira Interior
Covilhã, Portugal
jbfs@ubi.pt

***Abstract*—Insider threats remain a persistent cybersecurity challenge, leading to severe financial and reputational damage for organizations. Despite the widespread use of Role-Based Access Control (RBAC) and conventional monitoring tools, these measures often fail to detect sophisticated insider activities due to their reliance on static security controls [1]. Additionally, Insider Threat Analysis (ITA) is burdened by information overload, making detection highly dependent on human analysts [2]. To address these gaps, Snakey introduces a blue-team-focused keylogger designed for anomaly detection in keystroke behavior. This paper explores Snakey's architecture, security measures, and machine learning techniques, offering a scalable approach to proactive insider threat mitigation.**

I. Introduction

The increasing frequency of high-profile insider exploits [1] underscores the limitations of traditional insider threat detection strategies. Current Insider Threat Analysis (ITA) solutions suffer from information overload, making detection a human-intensive and error-prone process [2]. Research suggests two primary approaches to mitigate this issue: reducing the volume of data processed, or dividing the workload across multiple analysts. However, these solutions are insufficient for tackling threats that blend into normal user behavior.

Snakey seeks to bridge these gaps by introducing a keystroke-based anomaly detection system that leverages machine learning for behavioral analysis. The system captures and encrypts keystroke logs using AES-256-GCM [3], ensuring data confidentiality and integrity. By integrating with security dashboards like Splunk and Wazuh, Snakey enables real-time monitoring and proactive incident response. Furthermore, the use of Isolation Forests, an unsupervised machine learning technique [4], enhances anomaly detection without requiring large labeled datasets. This paper outlines Snakey's security architecture, detection methodologies, and compliance with NIST standards, positioning it as a valuable tool for mitigating insider threats in enterprise environments.

II. NIST CSF Integration and Implementation

The NIST Cybersecurity Framework (CSF) provides a structured approach to managing cybersecurity risks through five core functions: Identify, Protect, Detect, Respond, and Recover [5]. These functions help organizations establish a continuous security lifecycle, ensuring proactive threat mitigation and incident response. Snakey aligns with the NIST CSF by incorporating its principles into a keystroke-based anomaly detection system, enhancing enterprise security posture against insider threats. This structure is further detailed in figure 1.

A. Identify (Asset Management, Risk Assessment)

Snakey establishes baseline user behaviors to detect anomalies. It maps users to roles (Admins, Standard Users) and stores typical behaviors such as login times and typing habits, allowing it to identify deviations. Additionally, Snakey integrates real-time updates from threat intelligence platforms

such as AbuseIPDB and VirusTotal to assess the legitimacy of user logins. Privileged access is also monitored by tracking high-risk commands, such as unauthorized attempts to create new system users.

B. Protect (Access Control, Data Security, Safeguards)

To safeguard sensitive data, Snakey encrypts keystroke logs using AES-256 encryption [3] and secures key exchange through CRYSTALS-Kyber. The tool implements anti-tampering mechanisms, such as integrity hash checks, to prevent unauthorized modifications. Compliance with data privacy regulations is enforced through Role-Based Access Control (RBAC) and encryption standards, ensuring robust protection against unauthorized access.

C. Detect (Threat Detection, Anomaly Detection, Behavioral Analytics)

Snakey continuously analyzes keystroke behavior to identify threats. It detects abnormal typing speeds that may indicate automated brute-force attacks and monitors failed login attempts. Machine learning models, specifically Isolation Forests [4], profile user behavior to identify deviations indicative of insider threats. Real-time monitoring dashboards display alerts and provide SOC teams with live insights, integrating with Slack for immediate notifications.

D. Respond (Incident Response, Alerting, Automated Actions)

To streamline incident response, Snakey employs automated escalation based on risk scores. Suspicious behavior triggers alerts that escalate from simple warnings to account quarantines for high-risk activities. Adaptive authentication mechanisms, such as Time-based One-Time Passwords (TOTP), are enforced when anomalies are detected. Additionally, Snakey dynamically blocks malicious IPs and disables compromised accounts based on detected threats.

E. Recover (Post-Incident Analysis, Learning, and Refinement)

Following security incidents, Snakey enables post-incident forensics through encrypted log analysis. Its machine learning models undergo periodic retraining to adapt to emerging threats, ensuring continuous improvement. The system also reduces false positives through adaptive whitelisting, allowing legitimate user behaviors to be safely recognized over time.

III. Conclusions

Snakey represents a novel approach to detecting insider threats by combining traditional keylogging with advanced machine learning and encryption. By adhering to the NIST CSF, Snakey offers a structured solution for insider threat detection, ensuring data integrity, privacy, and compliance. Future enhancements will include continuous learning through machine learning and expanded integration with other security tools.

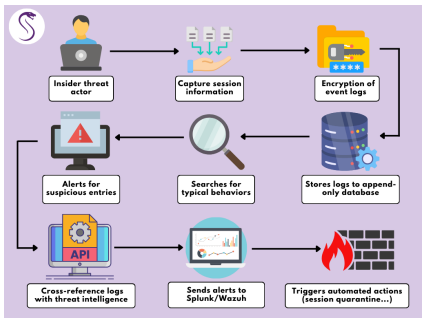


Fig. 1. Snakey System Architecture.

Acknowledgements

This work is funded by FCT/MECI through national funds and, when applicable, co-funded EU funds under UID/50008: Instituto de Telecomunicações.

References

[1] C. Probst, J. Hunker, and M. Bishop, *Insider threats in cybersecurity: The enemy within*. Springer, 2020.

[2] A. Chuvakin, *Security information and event management (SIEM) implementation*. Addison-Wesley, 2015.

[3] W. Stallings, *Cryptography and network security: Principles and practice*, 7th ed. Pearson, 2017.

[4] Y. Zhao, Z. Nasrullah, and M. Hryniewicki, "Pyod: A python toolbox for scalable outlier detection," *Journal of Machine Learning Research*, vol. 20, no. 96, pp. 1–7, 2019.

[5] National Institute of Standards and Technology, "Framework for improving critical infrastructure cybersecurity," https://www.nist.gov/cyberframework, Apr. 2023, accessed: 2025-04-04.