Adversarial AI attack detection: a novel approach using explainable AI and deception mechanisms

Maria NICULAE,
Beia Consult International, Bucharest, Romania
maria.niculae@beia.ro

George SUCIU,
Beia Consult International, Bucharest, Romania
george@beia.ro

Vlad STANESCU,
Beia Consult International, Bucharest, Romania
vlad.stanescu@beia.ro

Mari-Anais SACHIAN,
Beia Consult International, Bucharest, Romania
anais.sachian@beia.ro

Aristeidis FARAO,
University of Piraeus, Piraeus, Greece
arisfarao@unipi.gr

Athanasia SABAZIOTI,
University of Piraeus, Piraeus, Greece
a.sabazioti@unipi.gr

Christos XENAKIS,
University of Piraeus, Piraeus, Greece
xenakis@unipi.gr

Dionysios XENAKIS,
Department of Digital Industry Technologies of the National and Kapodistrian University of Athens, Athens, Greece
nio@uoa.gr

Ignacio LACALLE,
Universitat Politècnica de València, Valencia, Spain
iglaub@upv.es

Panagiotis Radoglou GRAMMATIKIS,
K3Y, Sofia, Bulgaria
pradoglou@k3y.bg

Nikolaos Sachpelidis BROZOS,
K3Y, Sofia, Bulgaria
nsachpelidis@k3y.bg

Zacharenia LEKKA,
K3Y, Sofia, Bulgaria
zlekka@k3y.bg

Giorgio BERNARDINETTI,
Consorzio Nazionale Interuniversitario per le Telecomunicazioni, Parma, Italy
giorgio.bernardinetti@cnit.it

Anastasia TSIOTA,
Fogus Innovations and Services, Athens, Greece
atsiota@fogus.gr

Georgios KALPAKTSOGLOU,
Fogus Innovations and Services, Athens, Greece
gkalpak@fogus.gr

Stylianos KARAGIANNIS,
PDM, Lisbon, Portugal
stylianos.karagiannis@pdmfc.com

Abstract
Detecting adversarial AI attacks has emerged as a critical issue since AI systems are becoming integral across all industries, from healthcare to finance and even transportation. Adversarial attacks stand on the fact that there exist weaknesses within machine learning and deep learning models, which they exploit on the grounds of their potential to cause serious disruptions and severe threats towards the integrity of AI operational procedures. In this light, the discussion will focus on developing robust mechanisms for detecting adversarial inputs in real-time to ensure that AI systems remain resilient against such sophisticated threats. While adversarial AI — software input sanitization, anomaly detection, and adversarial training — has some important foundational work, most approaches to them suffer from generalization challenges across attack types or real-time performance. This work will introduce novelty by extending the detection capabilities with explainable AI (XAI) and deception mechanisms. Adversarial activities will be detected based on adversarial training in combination with honeypots and digital twins, while keeping the process of detection transparent with XAI. While honeypots and digital twins decoy attackers, observing their behaviors can further strengthen detection methods. The results so-far promise tremendous improvements in the detection of adversarial attacks in high-risk AI applications, efficacy of honeypots for the capture of malicious behavior, and XAI for enhanced interpretability and reliability of the detection process. These techniques will enhance the robustness of AI systems against adversarial threats. Presented research contributes significantly by providing practical tools for cybersecurity professionals and AI practitioners against these attacks, thus offering new insights into AI for cybersecurity. The novelty value of the paper is the innovative integration of adversarial training, XAI, and deception techniques, which offers a combined, interpretable, and effective method toward the detection of adversarial AI attacks on cross-industry sectors.

Keywords: Adversarial AI detection, adversarial training, deception mechanisms, explainable AI, cybersecurity.

1. Introduction

The integration of Artificial Intelligence (AI) into critical sectors such as healthcare, transportation, and finance has brought transformative benefits but also significant vulnerabilities. Adversarial attacks, exploiting weaknesses in machine learning models, threaten the reliability and security of AI systems. This paper proposes a framework combining adversarial training, Explainable AI (XAI), and deception mechanisms to enhance adversarial attack detection and mitigation.

1.1. Background on adversarial AI threats

Following the increased use of artificial intelligence (AI) systems and machine learning (ML), there are also increased attempts to trick those AI systems. The techniques aiming to cause incorrect predictions or decisions by machine learning models are called adversarial attacks and have as their main goal to undermine trust and security in ML systems. Most of the time they do so using adversarial examples, which are perturbed inputs almost undetectable by humans. Adversarial attacks can be categorized based on either the attacker's knowledge or attack strategies.

Based on attacker's knowledge there are three types of attacks:

- White-box attacks, where the attacker has full knowledge of the model's architecture, parameters and training data. In a white-box attack the attacker uses this knowledge to create powerful adversarial examples able to lead the model into wrong decisions with high confidence. An example of a white-box attack is the Fast Gradient Sign Method (FGSM) that adds small perturbations to input images to maximize classification errors. In FGSM the attacker utilizes the sign of the gradient of the model's loss function to add the minimum possible perturbations capable of changing an input's predicted class. By using the loss function the attack is guaranteed to move the data point towards the right direction [1]. In [2], the authors used FGSM to check how adversarial attacks can trick detection systems that use AI into false alarms. Their main goal is to prove that AI can be used safely to identify potential anomalies and threats, in order to trigger the appropriate alarms in case of an attack.
- Black-box attacks, where the attacker does not have further information of the system, including the trained model, dataset, parameters, than a normal user, including the trained model, dataset, parameters. Black-box attacks are usually harder to succeed than other types of attacks because the attacker has no additional information than a normal user of the model. Also, due to the need for many more attempts to query the model, black-box attacks are more computationally expensive. An example of a black-box attack is the Boundary Attack that tries to find the decision boundary of a model by iteratively adding perturbation on an adversarial example. In this attack the attacker starts with a greatly perturbed input that is guaranteed to be misclassified. Then he proceeds to decrease the distance between the adversarial example and the original input until he reaches the minimum distance that tricks the model and remains undetected by humans [3].
- Grey-box attacks, where the attacker only has part of the available knowledge of a system, such as limited access to credentials or architecture details. These attacks combine aspects of black-box and white-box attacks in order to effectively trick the system. They are useful in many different cases, such as more realistic testing of defense systems when attacked by an insider who may know the model's architecture, but not the specific weights used in it. An example of a grey-box attack is the Gradient Estimation Attack that tries to approximate gradients by checking how the system responds to specific adversarial examples. This type of attack is limited because the attacker does not know the model's gradient or internal weights but has access to its input and output. Using this known information the attacker is reverse-engineering the model in order to find the missing information of the gradient [4].

Based on attack strategies there are four types of attacks:

- Evasion attacks, which are attacks designed to modify inputs to evade correct detection. An example of an evasion attack is to add undetectable by human perturbation in an image, so a stop sign is misclassified as a speed limit sign [5]. This type of attack regarding the traffic signs’ example is further analyzed in [6]. In this paper the authors propose their approach of defending against evasion attacks using an anomaly detector to find the potential anomalies and a reconstructor to create clean images of traffic signs.

- Poisoning attacks, where the attacker injects malicious or corrupted data into the training data sets, causing the AI model to produce inaccurate results. An example of poisoning attack is to add malicious samples to a dataset, so the model learns biased behaviors. For example, if a model used for animal recognition is trained with a dataset in which an adversary has added plenty of images where cats sit in baskets, the final model will misclassify a cat-input that does not sit in a basket [7].
- Model extraction attacks, where the attackers aim to reverse-engineer the model or steal its functionality by querying the system. An example of this type of attack is to copy an ML model by mimicking its behavior. In that case an attacker can provide specific input data to a model and collect the output in order to use it as label to train a different model that replicates the original [8].
- Inference attacks, where the attacker seeks to deduce sensitive information from the model. An example of an inference attack is when an attacker tries to determine whether a data point is part of the training set of the target model. In order to do this the attacker queries the model with input and then analyzes the confidence score of the result [9].

Below Fig. 1. illustrates the categorization of adversarial attacks, highlighting their classification based on two primary dimensions: the attacker's knowledge and the attack strategy employed.

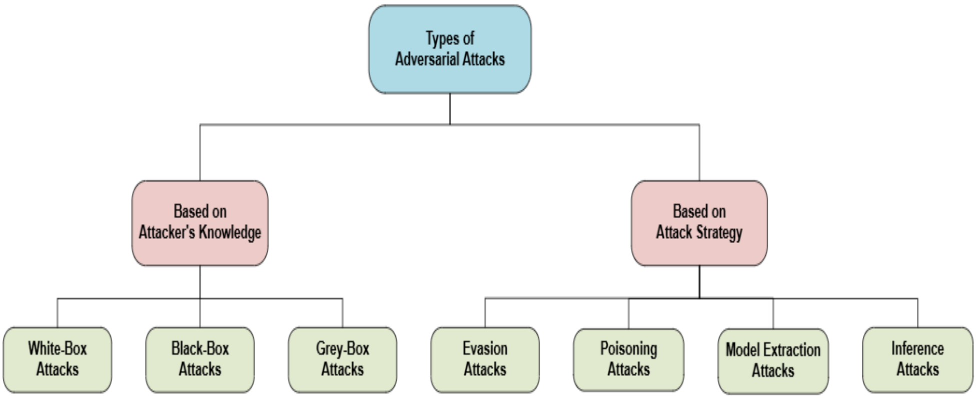

Fig. 1. Types of Adversarial Attacks

1.2. Importance of detecting adversarial AI Attacks

The increasing reliance on Artificial Intelligence (AI) across critical sectors such as healthcare, finance, and transportation makes these systems attractive targets for adversarial attacks [10]. These attacks exploit vulnerabilities in machine learning models, introducing subtle manipulations that often lead to incorrect outputs. Early and robust detection of such threats is crucial to ensure the safety, reliability, and trustworthiness of AI technologies, particularly in high-stakes environments.

Adversarial attacks have profound implications across multiple domains. For instance, in autonomous systems, adversarial inputs can compromise perception models, leading to catastrophic operational failures [11]. In security-sensitive applications like biometric authentication [12] and surveillance [13], such attacks can bypass detection mechanisms, enabling unauthorized access or evasion. Furthermore, adversarial attacks targeting financial systems can result in erroneous predictions, causing significant monetary losses [14]. Mitigating these risks requires not only detection mechanisms but also an understanding of adversarial behaviors to anticipate and counter evolving threats.

The significance of detecting adversarial attacks extends beyond immediate risk mitigation. It also strengthens the robustness of AI systems by exposing and addressing model vulnerabilities, enabling the development of more resilient architectures. This capability is vital to ensuring the secure and effective deployment of AI in increasingly complex and adversarial environments.

1.3. Objectives of the research

This research proposes a comprehensive framework for detecting adversarial AI attacks by combining adversarial training, Explainable AI (XAI), and deception mechanisms. The primary objectives include:

- Exploring advanced techniques for enhancing the detection capabilities of AI systems against a wide range of adversarial inputs. As adversarial attacks become increasingly sophisticated, they exploit subtle vulnerabilities in machine learning models to deceive them into making erroneous predictions. The research will be conducted for further developing methods that strengthen the robustness of models and their ability to generalize across attack types.
- Ensuring real-time detection of adversarial attacks is essential in dynamic and high-stakes environments where delays can have significant consequences, such as in autonomous vehicles or financial trading systems. Many existing methods struggle with real-time performance, limiting their effectiveness in dynamic environments where timely responses are crucial. This study will explore techniques that enable swift and accurate detection, ensuring that AI systems remain operational and secure during active threats [15].
- Integrating Explainable AI (XAI) techniques, such as Gradient-weighted Class Activation Mapping (Grad-CAM) [16] and SHapley Additive exPlanations (SHAP) [17], to provide stakeholders with clear, interpretable insights into detection outcomes. By making the detection process transparent, these techniques will foster trust among users and enable more informed decision-making. Additionally, the integration of explainability ensures that the framework's outputs are actionable, allowing cybersecurity professionals to validate and enhance system performance effectively.
- Leveraging Deception Mechanism such as honeypots and digital twins. Honeypots serve as a tool to lure and monitor adversaries, offering insights into their behavior and tactics in a controlled environment [18]. Digital twins, as virtual replicas of real systems, enable the safe simulation of adversarial scenarios without jeopardizing operational

infrastructure [19]. These mechanisms will not only strengthen the framework’s ability to anticipate and counter evolving threats but also contribute to a deeper understanding of adversarial techniques.

The remainder of this paper is structured as follows. In Section 2, we provide a comprehensive review of related work, covering (a) prior research in adversarial AI detection, (b) strategies and taxonomy of adversarial attacks, (c) gaps in existing solutions, and (d) advancements in Explainable AI (XAI) and deception mechanisms. Section 3 outlines the methodology adopted in this research, including (a) an overview of the proposed framework, (b) the integration of XAI techniques for interpretability, (c) approaches to adversarial AI detection, and (d) the roles of honeypots and digital twins in proactive defense. In Section 4, we detail the proposed framework, including its integration of adversarial training, XAI, and deception mechanisms, the conceptual model for real-time detection, and a comparative analysis with existing frameworks to highlight its advantages. Section 5 concludes the paper by summarizing the findings, discussing limitations, and outlining recommendations for future research to enhance the framework and address emerging adversarial challenges.

2. Related work

2.1. Prior research in adversarial AI detection

Adversarial attacks have emerged as a critical challenge in ensuring the robustness of machine learning models, particularly in fields like computer vision [20], natural language processing [21], and autonomous systems [22]. These attacks are characterized by subtle, carefully designed perturbations to input data, often imperceptible to human observers, that cause models to make incorrect predictions or classifications. The realization of such vulnerabilities has catalyzed a wave of research focused on understanding, detecting, and mitigating adversarial examples. Early investigations into these issues, such as those by Szegedy et al. (2014), revealed that even state-of-the-art deep neural networks (DNNs) were susceptible to targeted adversarial inputs, exposing a fundamental weakness in their design. This revelation underscored the urgent need for effective adversarial detection mechanisms to enhance the resilience of AI systems deployed in sensitive and high-stakes environments [23].

Recent advancements in adversarial AI detection have significantly expanded the range of techniques aimed at identifying and mitigating adversarial attacks across various domains. Modern approaches have moved beyond analyzing confidence scores and uncertainty metrics, as initially proposed by Hendrycks and Gimpel (2017), toward sophisticated frameworks and datasets that benchmark the performance of detection methods [24]. A notable contribution in this field is the ARIA dataset, introduced by Li et al. (2024). In their 2024 study, "The Adversarial AI-Art: Understanding, Generation, Detection, and Benchmarking," Li et al. delve into the complexities of AI-generated imagery, particularly focusing on adversarial scenarios. The researchers introduce the AdversaRIal AI-Art (ARIA) dataset, comprising over 140,000 images across categories such as artworks, social media visuals, news photographs, disaster scenes, and anime illustrations. This extensive dataset serves as a foundation for evaluating the effectiveness of both human and automated detection systems in distinguishing AI-generated images from authentic ones. Through comprehensive user studies and benchmarking of existing detection tools, the authors highlight the current challenges in accurately identifying AI-generated content, emphasizing the necessity for more advanced detection methodologies [25].

The development of auxiliary models or networks that operate alongside primary machine learning models continues to be a promising direction. These systems are trained to identify adversarial perturbations by analyzing specific features or activation patterns within the primary model. For instance, Metzen et al. (2017) demonstrated that integrating a small neural network into an existing model and using intermediate activations can effectively recognize adversarial inputs [26]. More recent approaches, such as Wang et al. (2023), introduce an innovative approach to detecting adversarial examples in deep neural networks (DNNs). The authors propose leveraging sentiment analysis techniques to identify adversarial perturbations by examining their progressive impact on the hidden-layer feature maps of DNNs. They design a modular embedding layer that transforms these feature maps into word vectors, assembling sentences suitable for sentiment analysis. Extensive experiments demonstrate that this detector surpasses existing algorithms in identifying recent attacks on models like ResNet and Inception, tested across datasets such as CIFAR-10, CIFAR-100, and SVHN. Notably, the detector comprises approximately 2 million parameters and can detect adversarial examples in under 4.6 milliseconds using a Tesla K80 GPU [27].

In the realm of text-based adversarial detection, Hu et al. (2023) introduced a novel framework designed to enhance the detection of AI-generated text. The RADAR framework employs adversarial training by integrating a paraphraser and a detector, where the paraphraser generates content aimed at evading detection, and the detector iteratively improves its ability to identify such content. This dynamic interplay enhances the detector's robustness against paraphrased AI-generated text. Evaluations across eight large language models, including LLaMA and Vicuna, demonstrate that RADAR significantly outperforms existing detection methods, particularly in scenarios involving paraphrased AI text. The study also highlights RADAR's strong transferability across different models, underscoring its potential for broad applicability in AI-text detection [28].

Despite these developments, challenges persist. The rapid evolution of adversarial attacks, such as mixed-prompt image-text generation, exposes vulnerabilities in existing detection systems. For example, Diao et al. (2024) examined the susceptibility of AI-generated image (AIGI) detectors to adversarial attacks. The authors introduce the Frequency-based Post-train Bayesian Attack (FPBA), a novel method that applies perturbations in the frequency domain to deceive AIGI detectors. FPBA leverages a post-training Bayesian strategy to enhance the transferability of adversarial examples across different model architectures, including Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs). Comprehensive experiments reveal that FPBA effectively compromises various state-of-the-art AIGI detectors, highlighting the pressing need for more robust detection mechanisms in the face of sophisticated adversarial threats [29].

The field of adversarial AI detection is progressing rapidly, propelled by the development of comprehensive datasets, novel frameworks, and hybrid methodologies. The ongoing dynamic between adversarial attackers and defenders highlights the imperative for sustained innovation and interdisciplinary collaboration to safeguard the security and reliability of AI systems, particularly those deployed in critical domains.

2.2. Gaps in existing solutions

In the context of advancing cybersecurity measures, artificial intelligence (AI) has emerged as a pivotal tool in addressing contemporary challenges, including the detection of advanced threats and mitigation of complex cyberattacks. As highlighted in the study AI & Cybersecurity, AI technologies facilitate rapid anomaly detection, automate incident response processes, and enable the development of proactive systems that significantly reduce reaction times to emerging threats. These advancements underline the importance of integrating AI-driven methodologies into adversarial defense frameworks to enhance robustness and adaptability in real-world scenarios [30].

A recent study highlights how the rapid digitization of institutions, accelerated by events like the COVID-19 pandemic, has exposed significant vulnerabilities in data security frameworks. This transition, often implemented without rigorous testing, has led to increased exposure to cyberattacks exploiting human error and weak digital infrastructures. The work of Popa (2024) particularly underscores the dual-edged role of AI in these scenarios, where attackers increasingly use AI to craft sophisticated, targeted attacks, such as phishing and malware, while defenders struggle to adapt security measures to this evolving threat landscape. Addressing these gaps requires a concerted effort to integrate AI-driven defense mechanisms, emphasizing proactive measures like risk assessment, data encryption, and two-factor authentication, as highlighted in the study. These challenges illustrate the urgent need for comprehensive, adaptive solutions to safeguard critical data systems in an increasingly digitized world [31].

Despite the significant progress in adversarial defense mechanisms, numerous critical gaps persist, limiting the robustness and reliability of machine learning models in adversarial settings. Addressing these gaps is essential for the secure and effective deployment of AI systems in real-world applications.

Many existing defense strategies are tailored to specific adversarial attacks, resulting in limited effectiveness against novel or slightly modified attack methods. This specificity fosters a reactive cycle, wherein defense mechanisms lag behind the rapidly evolving tactics of adversaries, leaving AI systems vulnerable to unforeseen threats [32].

Adversarial examples crafted to deceive a particular model often succeed against other models, even those with different architectures or training data—a phenomenon known as transferability. This issue is particularly problematic in black-box attack scenarios, where attackers have limited knowledge of the target model. The transferability of adversarial examples enables attackers to exploit vulnerabilities across multiple models, thereby diminishing the effectiveness of defenses that focus solely on direct attack prevention [33].

Enhancing a model's resilience to adversarial attacks often incurs increased computational complexity and reduced accuracy on benign inputs. This trade-off poses significant challenges, especially for resource-constrained applications that demand both high performance and robust defenses. Achieving an optimal balance between robustness and performance remains an unresolved issue in adversarial defense [34].

Adversaries continually develop innovative methods, such as employing generative models to craft more sophisticated adversarial examples. This dynamic landscape necessitates proactive defense mechanisms capable of addressing current threats while anticipating and adapting to future attack strategies. The evolving nature of adversarial tactics complicates the development of static defense solutions [35].

Many proposed defense solutions are evaluated under controlled, theoretical conditions that do not adequately reflect the complexities of real-world environments, including environmental noise, sensor imperfections, or hardware constraints. This disparity can lead to overestimated defense performance and unprepared systems in operational contexts, highlighting the need for defenses that are effective in practical applications [36].

Numerous adversarial defenses function as "black-box" solutions, obscuring their decision-making processes. This opacity hinders the identification of potential weaknesses and impedes the development of more robust and explainable solutions. Improving the interpretability and transparency of defense mechanisms is crucial for building trust and facilitating the advancement of effective defenses [32].

The presence of bias in artificial intelligence systems remains a significant challenge in ensuring fairness and equity across diverse applications. As highlighted by Boce (2022), biases, such as racial and gender biases, are often embedded in machine learning algorithms through the training data or modeling techniques used. These biases can lead to unfair outcomes, such as disproportionate risk assessments or discriminatory decisions in areas like criminal justice and hiring practices. Addressing such biases is crucial for developing equitable AI systems. Boce’s analysis emphasizes the importance of strategies such as pre-processing data, implementing counterfactual fairness methods, and using decoupled classifiers to mitigate discrepancies and enhance fairness in AI decision-making [37].

The lack of universally accepted protocols for assessing defense strategies hampers objective comparison and measurement of progress in the field. This absence of standardization impedes the development of effective, broadly applicable adversarial defenses, underscoring the necessity for establishing standardized evaluation frameworks [20].

To address these challenges, future research must focus on developing generalized and proactive defense mechanisms, enhancing the interpretability of models, and establishing standardized evaluation protocols. These efforts are critical to bolstering the robustness, reliability, and real-world applicability of AI systems in the face of sophisticated adversarial threats.

2.3. Advancements in explainable AI and deception mechanisms

In the rapidly evolving landscape of AI and cybersecurity, Waizel (2024) provides a comprehensive examination of the interplay between AI-driven cyberattacks and corresponding AI-based defensive measures. The article underscores the sophistication of AI-enabled offensive strategies, such as machine learning-based malware and AI-generated phishing schemes, and juxtaposes these with state-of-the-art defensive mechanisms, including anomaly detection and automated threat responses. This nuanced analysis reveals significant gaps in the current defensive capabilities, emphasizing the urgent need for adaptive and anticipatory security frameworks. These insights align with the objectives of this study, which seeks to integrate advanced techniques like deception mechanisms and Explainable AI to bridge the existing disparities in AI cybersecurity dynamics [38].

Recent advancements in Explainable AI (XAI) have significantly transformed adversarial detection by providing transparency in decision-making processes. Techniques such as Local Interpretable Model-agnostic Explanations (LIME) [39] and SHapley Additive exPlanations (SHAP) [40] have made it possible to understand the influence of specific input features on model outputs, offering valuable insights into adversarial manipulations. For instance, Grad-CAM (Gradient-weighted Class Activation Mapping) [41] visualizes the regions of input data that drive model predictions, enabling researchers to identify anomalies introduced by adversarial attacks.

These advancements have been instrumental in addressing the "black-box" nature of machine learning models, particularly in adversarial contexts. However, challenges remain, including the computational demands of XAI methods, which can hinder their scalability in real-time applications, and the vulnerability of XAI tools to adversarial attacks targeting interpretability frameworks. Research is increasingly focusing on developing lightweight, adversary-resilient XAI methods that balance interpretability with efficiency, ensuring their applicability in operational systems [42].

Parallel to advancements in XAI, deception mechanisms such as honeypots and digital twins have emerged as proactive strategies in adversarial detection. Honeypots, adapted from traditional cybersecurity applications, act as decoy environments designed to attract and engage adversaries. These systems collect data on adversarial behaviors and attack methodologies, offering critical insights into how adversarial examples are generated and deployed [43]. Recent innovations in dynamic honeypots, which evolve in response to adversarial activity, have enhanced their efficacy in capturing real-time attack patterns while minimizing the risk to operational systems [44].

Digital twins provide a complementary approach by creating virtual replicas of real-world systems that can simulate adversarial scenarios in a controlled and risk-free environment. These simulations allow researchers to evaluate the impact of adversarial attacks and to test the robustness of detection and defense mechanisms without compromising live systems [45]. Recent advancements in digital twin technology include the integration of real-time data and predictive analytics, enabling more accurate modeling of adversarial behaviors and adaptive responses [46]. By incorporating digital twins into adversarial detection frameworks, organizations can proactively identify vulnerabilities and refine defense strategies before adversaries exploit them.

Despite the progress in XAI and deception mechanisms, significant challenges remain. XAI methods often require computationally intensive processes, limiting their applicability in real-time detection systems. Additionally, the design of effective honeypots and digital twins must carefully balance the collection of adversarial data with the risk of inadvertently providing adversaries with information that could aid future attacks. Addressing these challenges requires continued innovation in algorithmic efficiency, system integration, and adversary-aware design principles.

3. Methodology

3.1. Framework overview

The proposed framework for adversarial AI detection is designed to address critical challenges in safeguarding AI systems from adversarial attacks. By integrating multiple advanced techniques, the framework aims to enhance robustness, interpretability, and adaptability in high-stakes environments such as healthcare, finance, and autonomous systems. The framework comprises three primary components: detection, explanation, and response, which work in unison to identify, interpret, and mitigate adversarial threats effectively.

At the core of the proposed framework is a detection module that will leverage adversarial training and anomaly detection techniques to identify malicious inputs. The envisioned module plans to combine statistical anomaly detection with adversarially robust neural network architectures. Adversarial training will aim to improve the model's ability to generalize across diverse attack types by exposing it to adversarial examples during the training phase. Simultaneously, an anomaly detection layer is intended to evaluate input features for irregularities, flagging data that deviates from normal distributions. This dual-layer approach is expected to provide a comprehensive and reliable detection mechanism.

To address the "black-box" nature of AI systems, the framework will integrate an explainability module powered by Explainable AI (XAI) techniques [47]. This module is designed to provide transparent and interpretable insights into the detection process by highlighting input features that contribute to model predictions. Methods such as Gradient-weighted Class Activation Mapping (Grad-CAM), SHapley Additive exPlanations (SHAP), and Local Interpretable Model-agnostic Explanations (LIME) are expected to be employed to generate visual and textual explanations. These explanations aim to enable stakeholders, including cybersecurity experts and system operators, to understand the nature of adversarial attacks and validate detection outcomes.

Beyond detection and interpretation, the framework will include a decision and response module designed to mitigate detected threats. This module is conceptualized to use game-theoretic approaches and adaptive defense strategies to recommend optimal countermeasures. Depending on the threat's nature and severity, the module could suggest actions such as input rejection, reclassification, or reconfiguration of system parameters. By balancing automated responses with actionable insights for human operators, the envisioned module aims to enable swift and effective responses to adversarial attacks while maintaining operational continuity.

The proposed framework is designed to operate as a unified system where the detection module identifies potential threats, the explainability module interprets and validates detection outcomes, and the decision and response module implements or recommends appropriate countermeasures. Each component is intended to function autonomously while integrating seamlessly into a cohesive workflow. This integration is expected to ensure that adversarial activities are not only detected but also understood and addressed in a timely and effective manner.

3.2. Explainable AI (XAI) integration

Explainable AI (XAI) plays a crucial role in the proposed framework by addressing the opacity often associated with machine learning models, particularly in adversarial detection contexts. By integrating XAI techniques, the framework aims to make the detection process transparent and interpretable, enabling stakeholders to better understand how decisions are made and how adversarial attacks affect model behavior. By employing tools such as SHAP and Grad-CAM, the framework provides interpretable visualizations and explanations that make it possible to understand how adversarial inputs influence model decisions. For example, Grad-CAM highlights regions of input data affected by adversarial perturbations, allowing stakeholders to identify the pathways through which attacks exploit model vulnerabilities [48].

To deepen the understanding of adversarial attacks, the framework will integrate counterfactual analysis as part of the XAI module. Counterfactual explanations identify minimal changes to inputs that would alter model predictions, offering a clear view of how adversarial manipulations deviate from normal decision boundaries. This approach not only aids in detecting adversarial examples but also helps refine the boundaries of the detection framework, making it more resilient to new attack strategies.

Adversarial attacks can also target the interpretability of detection systems, attempting to obscure malicious behavior by manipulating explanation outputs. To address this, the framework proposes to develop adversary-resilient XAI techniques. These methods will include noise-resistant feature attribution and perturbation-aware explanations, ensuring that the interpretability module remains reliable even under adversarial conditions.

The XAI integration will also emphasize usability by providing interactive visualization tools for cybersecurity professionals and system operators. These tools will offer clear visual overlays, sensitivity maps, and anomaly detection metrics, enabling users to explore adversarial activities and validate detection outcomes. The interactive nature of these tools will facilitate informed decision-making and collaborative analysis in real-world scenarios.

By incorporating XAI into the proposed framework, this research aims to bridge the gap between technical robustness and operational usability in adversarial detection. Explainability will enhance trust in AI systems, enable stakeholders to validate and act on detection outcomes, and provide insights that can guide the iterative improvement of detection mechanisms. The integration of XAI represents a critical step toward creating transparent, reliable, and secure AI systems capable of addressing the complexities of adversarial attacks.

3.3. Adversarial AI detection

Adversarial AI involves intentionally altering inputs to take advantage of weaknesses in ML models, typically aiming to confuse or undermine these systems. This method usually entails introducing slight, barely noticeable changes to input data—like tweaking an image or making minor adjustments to text—leading AI models to generate inaccurate results [49]. For instance, an image classifier could confuse a slightly modified stop sign with a speed limit sign, which could result in serious consequences for autonomous driving systems. The stakes are significant in areas such as cybersecurity, fraud detection, medical diagnosis, and autonomous systems, where the precision and dependability of AI are crucial. The *Adversarial AI Detection* tool aims to tackle these challenges by recognizing altered inputs, protecting AI systems, and reducing risks linked to adversarial attacks [50], thereby fostering trust and resilience in their functioning.

The *Adversarial AI Detection* tool uses a comprehensive strategy, blending sophisticated ML methods with specialized knowledge from the field. At first, it employs statistical and ML techniques to examine input data for any unusual patterns. For instance, it can examine aspects like unusual pixel distributions, noise patterns, or semantic inconsistencies to identify possible adversarial interference. Furthermore, the tool includes adversarial training, an approach where the AI model encounters challenging examples throughout its training process [51]. By recognizing these manipulations, the model builds a greater resilience to attacks in real-world situations. Additionally, the tool utilizes methods to monitor how particular input features affect model predictions. These methods assist in identifying inconsistencies [52] that could indicate potential tampering [53], offering a clear and understandable framework for both operators and developers.

To develop the *Adversarial AI detection tool*, it is needed to integrate it with existing AI systems. This integration aims to enhance security and minimize interruptions at the same time. Such a tool will operate in real time to accomplish essential tasks, including but not limited to detecting malicious activities (e.g., in financial transactions) or overseeing inputs (e.g., in autonomous systems) swiftly identifying and correcting any potentially malicious inputs [54]. These activities are carried out to meet the needs of the AI-based system. In situations where time is not a pressing concern, it operates in batch processing to review data from the past, guaranteeing a comprehensive analysis. The tool is fundamentally structured for life-long enhancement. Input from recognized adversarial ML is integrated into the model's learning process, enhancing its ability to recognize new attack methods [55]. The system draws on insights from wider networks, keeping itself informed about new tactics used by adversaries. The tool, along with regular testing against new adversarial examples, guarantees lasting strength, adapting as it faces increasingly complex threats.

3.4. Role of honeypots and digital twins

Honeypots are cybersecurity tools that decoy a real system without exposing real software systems behind with the intention of disguising an attackable point with the final goal of deriving insights of potential hackers. The ultimate goal of honeypots is multi-dimensional, ranging from generating statistics to allowing the creation of counter-measures or specific policies [56] that will enhance the cyber-protection of (critical, or not) systems. They also serve the purpose of acting as early alarms to prevent the system administrators and to increase awareness of the vulnerabilities sought by broad attackers. Honeypots can increase an infrastructure's security as they: a) consume the attacker's resources and b) record and analyse the attacker's actions [57].

The use of deception technology, particularly modern honeypots, has proven effective in mitigating cyber threats in smart city infrastructures. Waizel (2022) highlights the utility of honeypots in providing early detection capabilities against Advanced Persistent Threats (APTs) and ransomware attacks. By leveraging deception technology, security teams can bait attackers, delay their actions, and gather critical intelligence on attack methodologies. The proposed modern honeypot triangle model integrates high-interaction decoys and dynamic lures, offering robust protection for critical systems while enabling proactive threat remediation. These insights emphasize the strategic role of deception mechanisms in enhancing the resilience of smart city systems against sophisticated cyber threats [58].

Honeypots have appeared in the literature since late 2000s [59, 60], where their nature and growing interest were already discussed. Since then, several studies emanate a potential characterization of those, mostly from the viewpoint of coverage depth [61, 62, 63].

- **Low-Interaction Honeypot**: Disguising endpoints that will allow little penetration by an attacker. Building on sniffing, packet analysis and basic inspection, these collect information such as attacker origin, frequency and gauging which types of attacks are more prone to be used [64].
- **Mid-Interaction Honeypots**: More sophisticated systems that allow a certain degree of interaction, aiming at providing an impression of partial success to the attackers. However, they are not equipped with depth, and are limited to the different basic characteristics of specific underlying disguised services (e.g., network [65], remote log via SSH [66]...), per type of attacks (DoS [67], ransomware [68], SQL injection [69]) or per target deployments (e.g., IoT-networks [67], cyber-physical systems [70]...).
- **High-Interaction Honeypots**: Fully-functional systems that may act as replicas (or Digital Twins - see later) of a real underlying infrastructure, oriented to gather advanced information about attackers, such as behaviour or trends [71]. Latest research on this type of honeypots builds on AI and is providing very promising results [72, 73, 74].

In the environment of interest, several open source honeypots are proposed to take part. Specific tools such as *cowrie* [75], Kippo [76], *dionaea* [77], YAKSA [78] or T-POT [79] are already being integrated. As per the on-going implementation, the design consists of including such elements on-demand adjusted to user requirements.

In the designed framework, monitoring and sniffing tools as well as high-interaction honeypots will collect adversarial activity over so-called virtual objects (that are a virtual representation of a real asset of company existing in its ICT systems), Big data management, data harmonization and data correlation can be utilised in the form of honeypots to extract knowledge from adversarial activity in the “deceptive” simulated environment to enhance the detection and response capabilities of the “real” system.

The role of such compacted provision is to imitate the whole organization (e.g., infrastructure, systems, employees, etc.) in order to trick adversaries and collect the information generated by the attacks. The attacking entities are led to believe that they are attempting to compromise the actual infrastructure and not a simulated environment. This is particularly true when honeypot data are correlated on a SIEM tool, minimizing threat analyst efforts. For the former case, the use of game theoretic strategies [80] enable honeypots to efficiently deceive and delay potential attackers.

As a matter of fact, the combination of honeypots with other cyber-security resources has proven usual in the literature in the past few years. High-interaction honeypots have been paired in the past with artificially created personas (virtual personas) to become a better decoy [81].

One of the mechanisms utilized in the design framework is, precisely, the combination of honeypots with Digital Twins. Digital Twins are real-time, digital replicas of physical objects, systems, and processes. They are designed to mirror the functions and operations of their real-world counterparts, with continuous updates from sensors or other data sources (e.g., network traffic) to accurately reflect the evolving state and behavior of the physical entity. Digital Twins are continuously updated with real-time data, enabling faster market entry, cost savings, and risk reduction. Used for products, machinery, or entire business ecosystems, it offers insights into past performance, optimizes current operations, and predicts future outcomes. Digital twins are unique to their physical counterparts, simulate behavior with high fidelity, and update in real-time (potentially using IoT sensors, probes, or other communication technologies). They enable physical-digital convergence by mirroring changes between the twin and its counterpart.

In the context of cybersecurity, Digital Twins are widely used, as depicted in the thorough literature review by Pokhrel, Katta, and Colomo-Palacios [82]. Digital Twins are used, for instance, to replicate firewall features, prevent incidents, replicate attack generators (or simulate attacks), to create training playgrounds for companies without facing any risks, for penetration testing or for assessment of vulnerabilities, among others.

Digital Twins have recently seen a great advancement and they have been applied in several domains, such as IoT, 5G, healthcare, automotive, etc [83]. However, the combination of Digital Twins with cybersecurity tools, such as honeypots has not been thoroughly studied yet [84].

In this work, authors propose a scenario where events are thrown into a Digital Twin replica of complex systems, and information must be extracted. Events range from legit requests to diverse types of cyberattacks such as Denial of Service, privilege escalation or SQL injection. Also, digging deep into the greenfield of adversarial AI attacks, the honeypot and Digital Twin will be exposed to synthetically-generated data poisoning, evasion and transfer attacks. All the previous is being tested on different business cases that aim at representing the heterogeneity of entities that are targeted by mentioned attacks (for example, industrial SMEs handling robotic arms, technological SMEs offering digital services, health companies…). Events will be registered and handled by network monitoring mechanisms embedded within the honeypots. The honeypots will be, in turn, disguising functionalities and services building on top surreal systems. Indeed, Digital Twins will lie behind, ultimately protecting real systems while allowing accurate traceability and study. Such honeypots will diversify the response to the generated attacks, and, after a proper connection to security analytics module that will report all gathered knowledge into the Adversarial AI engine discussed in previous sections.

Remarkably, in the proposed framework (see later), authors go beyond the state of the art by investigating the combination of intelligent deception technologies such as high interaction honeypots, digital twins, and virtual personas to create a novel high-interaction deception layer that will deceive attackers by posing as an easy (i.e., vulnerable) target. Thus, the usage in the same environment of honeypots (such as based on T-POT) with Digital Twins of full-fledged IT systems (for instance, a company hosting cloud services, a Wordpress website, a mail server and a remote connection service) will provide a thorough and realistic representation of the organization environment that not only will conceal their presence from fingerprinting attacks, but also it will allow the collection of rich data of adversarial activity, which can be ingested by AI-powered modules improving the company's resilience to Adversarial AI attacks.

Fig. 2 illustrates the architecture of the deception layer within the proposed framework, designed to engage, analyze, and respond to adversarial threats.

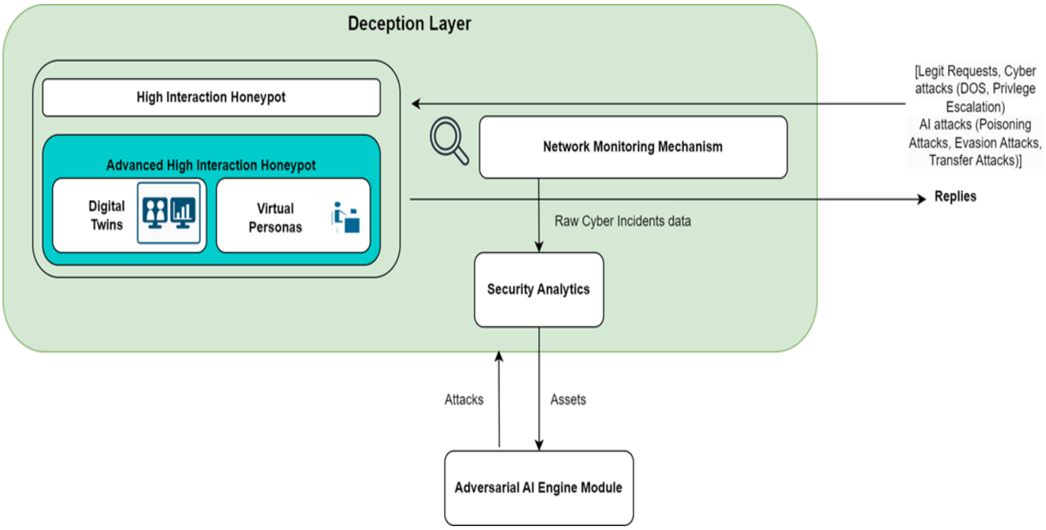
Deception Layer
High Interaction Honeypot
Advanced High Interaction Honeypot
Digital Twins
Virtual Personas
Network Monitoring Mechanism
[Legit Requests, Cyber attacks (DOS, Privlege Escalation)
AI attacks (Poisoning Attacks, Evasion Attacks, Transfer Attacks)]
Replies
Raw Cyber Incidents data
Security Analytics
Attacks
Assets
Adversarial AI Engine Module

Fig. 2. Deception Layer.

In addition, framework's deception layer will accommodate novel AI-based deception and monitoring tools (including time regression, reinforcement learning, among others) aiming at the deception of the attacker and the extraction of knowledge to empower organisations with the ability to predict and respond to future cyberattacks. These blocks, and the whole architecture, are being designed to align with the required Digital Twin Platform Stack [85] defined by the Digital Twin Consortium.

4. Proposed framework

4.1. Integration of adversarial training, XAI, and deception mechanisms

The proposed framework integrates adversarial training, Explainable AI (XAI), and deception mechanisms into a cohesive system to address the limitations of existing adversarial detection methods. This integration is designed to enhance the robustness, interpretability, and adaptability of AI systems operating in adversarial environments. Each component of the framework contributes a unique capability, ensuring a comprehensive approach to adversarial detection and mitigation.

Adversarial training forms the foundation of the framework by strengthening model robustness against adversarial examples. This process involves exposing the model to crafted adversarial inputs during the training phase, allowing it to learn patterns that differentiate normal data from malicious perturbations. The training procedure includes diverse attack scenarios, ensuring the model's generalization across various attack types. By iteratively refining the model with adversarial examples, the system becomes more resilient to evasion tactics commonly used in adversarial attacks. However, adversarial training alone can be computationally expensive and may not generalize well to unknown attack vectors, necessitating the integration of complementary techniques.

To address the interpretability gap in traditional adversarial detection systems, the framework incorporates XAI techniques that provide transparent and actionable insights into detection outcomes. The use of tools such as Grad-CAM, SHAP, and LIME allows for the decomposition of model predictions into human-understandable explanations. These methods highlight the features that influence model decisions, making it possible to identify the specific pathways through which adversarial perturbations manipulate model behavior.

Deception mechanisms, including honeypots and digital twins, complement adversarial training and XAI by providing proactive defense strategies. Honeypots are strategically deployed decoy systems designed to imitate real-world infrastructure, luring attackers into a simulated environment while safeguarding actual systems. By attracting adversaries, honeypots gather detailed insights into attack vectors, behavior, and methodologies. This data is instrumental for refining detection algorithms and devising countermeasures. In the proposed framework, high-interaction honeypots, capable of simulating fully functional systems, play a pivotal role. These honeypots act as replicas of operational IT environments, mimicking services and interfaces to convincingly deceive attackers. Moreover, they consume the attacker's resources, delay progression, and increase the likelihood of detection. Digital twins provide a complementary layer of deception by replicating physical systems, networks, or processes in real-time digital simulations. Unlike honeypots, which primarily serve as traps, digital twins enable the safe simulation of adversarial scenarios, allowing the framework to test and refine its defenses without compromising operational systems. The proposed framework employs digital twins to mirror complex organizational infrastructures, such as cloud services, IoT systems, and industrial control environments. These replicas continuously update based on real-time data, ensuring they remain accurate representations of their physical counterparts. In the adversarial AI context, digital twins are exposed to synthetic attacks to analyze their impacts and refine detection strategies. This simulation environment enables the framework to diversify its responses to adversarial behaviors, enhancing its adaptability to evolving threats. Additionally, the integration of digital twins with AI-powered analytics modules ensures that insights derived from simulated attacks contribute to improving the framework's overall resilience. The synergy between honeypots and digital twins creates a robust deception layer within the framework. While honeypots actively engage adversaries and collect detailed behavioral data, digital twins provide a controlled environment for testing and refining defensive strategies. Together, they form a comprehensive mechanism that not only deceives and delays attackers but also generates actionable intelligence to strengthen AI-powered adversarial defenses. The proposed framework advances the state of the art by integrating these technologies into a unified deception layer. By combining high-interaction honeypots with real-time digital twins and virtual personas, the framework creates a realistic, high-fidelity environment that convincingly mimics operational systems.

The integration of these components ensures that the proposed framework is not only robust against adversarial attacks but also interpretable and adaptive to evolving threats. Adversarial training strengthens the model's foundational resilience, while XAI techniques ensure that detection processes remain transparent and actionable. Deception mechanisms offer a proactive approach to understanding and mitigating adversarial activities, making the framework adaptable to a wide range of application domains and attack scenarios. This multifaceted integration represents a significant advancement in adversarial AI detection, bridging critical gaps in current methodologies and paving the way for secure and reliable AI systems.

4.2. Conceptual model for real-time detection

The proposed framework includes a conceptual model for real-time detection, designed to ensure timely identification and mitigation of adversarial threats. Real-time detection is critical for AI systems operating in dynamic and high-stakes environments, where delays in recognizing adversarial attacks can lead to significant disruptions or harm. The model combines lightweight anomaly detection techniques, adaptive response mechanisms, and Explainable AI (XAI) tools to achieve both speed and accuracy in detection.

At the core of the conceptual model is a lightweight anomaly detection pipeline that evaluates input data for deviations from expected patterns. This pipeline incorporates both statistical methods and neural network-based approaches to flag inputs that exhibit characteristics associated with adversarial perturbations. The use of lightweight techniques ensures that the detection process imposes minimal computational overhead, allowing the system to maintain real-time performance without compromising its ability to identify diverse attack types. To further enhance reliability, the anomaly detection module integrates with adversarially trained neural networks, which are preconditioned to resist common adversarial tactics.

A key component of the real-time detection model is the adaptive thresholding mechanism, which dynamically adjusts sensitivity based on the nature of incoming data and the operational context. This adaptive approach reduces the likelihood of false positives, ensuring that only genuine threats are flagged for further action. By tailoring detection thresholds to the system's environment, the model remains both efficient and accurate across a variety of use cases, including autonomous systems, financial applications, and healthcare diagnostics.

To support real-time decision-making, the model incorporates Explainable AI techniques that generate interpretable insights into detection outcomes. Gradient-based methods such as Grad-CAM are employed to provide visual explanations, highlighting regions of the input data most affected by adversarial perturbations. These explanations are generated in near real-time, enabling operators to understand the nature of the threat and validate the system's response. By making detection processes transparent, the XAI component builds trust and facilitates collaboration between automated systems and human operators.

The real-time detection model also integrates seamlessly with the framework's decision and response module, which leverages game-theoretic approaches to recommend optimal countermeasures. Upon detecting a threat, the module assesses the severity and context of the attack to suggest actions such as input rejection, reclassification, or reconfiguration of system parameters. This integration ensures that responses are both swift and contextually appropriate, minimizing the impact of adversarial activities while maintaining operational continuity.

The conceptual model emphasizes scalability and adaptability, making it suitable for deployment across a wide range of domains. Its modular design allows for the inclusion of domain-specific optimizations, such as incorporating additional detection techniques for specific types of adversarial attacks. Furthermore, the lightweight nature of the detection pipeline ensures that the model can operate effectively in resource-constrained environments without sacrificing performance.

4.3. Comparative analysis with existing frameworks

In recent years, the field of adversarial AI detection has seen significant advancements, with various frameworks proposed to enhance the robustness and security of AI systems. A comprehensive review published in the Journal of Big Data in August 2024 analyzed over sixty recent studies on AI-driven detection and prevention of cyber-attacks, highlighting the rapid increase in the number and sophistication of such attacks and stressing the importance of developing effective detection and prevention strategies to mitigate potential damages.

Despite these advancements, existing frameworks often face challenges in scalability, particularly when dealing with high-dimensional or multi-modal data. For example, frameworks optimized for image data may struggle to handle text or combined image-text inputs effectively. The proposed framework is designed with scalability in mind, incorporating modular components that can be customized for specific domains. Multi-modal XAI techniques allow the framework to adapt to complex data types, while the lightweight nature of the detection pipeline ensures scalability in resource-constrained environments. This adaptability makes the framework suitable for deployment across diverse applications, from autonomous systems to financial fraud detection.

Few existing frameworks prioritize real-time detection, often sacrificing speed for accuracy or robustness. This trade-off limits their applicability in time-sensitive environments, such as autonomous vehicles or financial trading systems. The proposed framework incorporates lightweight anomaly detection techniques and efficient XAI methods to ensure real-time performance. Features such as adaptive thresholding and low-latency explainability enable the system to operate effectively under dynamic conditions, delivering timely and accurate detection outcomes without significant computational overhead.

Another area where the proposed framework differentiates itself is the integration of deception mechanisms. These mechanisms are relatively underexplored in existing frameworks but offer significant potential for proactive defense. Honeypots provide valuable data on adversarial behaviors by engaging attackers in decoy environments, while digital twins simulate adversarial scenarios in a controlled setting, enabling the refinement of detection strategies. By incorporating these components, the proposed framework not only enhances its robustness but also anticipates and mitigates emerging threats.

The proposed framework surpasses existing approaches by offering a holistic solution that combines robustness, interpretability, scalability, and real-time performance. Its integration of deception mechanisms and advanced XAI techniques ensures a proactive and transparent approach to adversarial detection, addressing key limitations in current methodologies. By bridging these gaps, the framework aims to set a new standard for adversarial AI detection systems, enabling secure and reliable operation in high-stakes environments.

5. Conclusion and future directions

5.1. Summary of findings

This study introduced a novel framework for detecting and mitigating adversarial AI attacks by integrating adversarial training, Explainable AI (XAI), and deception mechanisms. The proposed framework addresses key challenges in adversarial detection, including enhancing robustness, improving real-time detection capabilities, and ensuring transparency through explainable methodologies. The use of deception mechanisms, such as honeypots and digital twins, adds a proactive layer to detect and understand adversarial behaviors. By combining these elements, the framework demonstrates potential to advance the resilience and adaptability of AI systems across diverse high-stakes applications.

The findings emphasize that integrating these components creates a more comprehensive and dynamic defense strategy. Adversarial training enhances model robustness against diverse attack types, XAI ensures interpretability and trust, and deception mechanisms contribute to proactive monitoring and threat analysis. Collectively, these techniques advance the state of the art in adversarial AI defense by bridging critical gaps identified in existing solutions.

5.2. Recommendations for future research

While the proposed framework demonstrates significant advancements, this research has focused on conceptual development, with implementation planned as the next phase. Future exploration is needed to refine and operationalize the framework, addressing several key areas to enhance its adaptability, scalability, and practical application. A critical direction involves improving the framework's dynamic adaptation capabilities to address the evolving nature of adversarial techniques. Threats leveraging generative AI or hybrid attack strategies continue to pose sophisticated challenges, and ensuring the framework's effectiveness will require further research into proactive and anticipatory detection methods.

Expanding testing to include large-scale, real-world environments will be vital for validating the framework's scalability and performance under diverse operational conditions. Real-world deployments often introduce complexities, such as environmental variability and domain-specific constraints, which cannot be fully captured in controlled experimental setups. Future research should focus on bridging this gap by simulating or deploying the framework in practical scenarios across domains like healthcare, finance, and transportation. Additionally, collaboration with industry stakeholders to establish standardized benchmarks and evaluation protocols will facilitate consistent comparison and provide a clear measure of progress in adversarial defense research.

Another area requiring further exploration is the development of lightweight Explainable AI (XAI) methods that prioritize computational efficiency without sacrificing interpretability. Current XAI techniques, while powerful, often impose significant computational demands, limiting their applicability in resource-constrained settings such as edge devices or real-time systems. Research should aim to optimize these methods for broader adoption, ensuring transparency while maintaining the performance required for operational systems.

By building upon the insights gained from this research, future work can contribute significantly to advancing the field of adversarial AI defense, ensuring secure and resilient AI systems in increasingly complex and adversarial environments.

Acknowledgements

This research has received funding from European Commission's Horizon Europe research and innovation program under grant agreement No. 101131292 (AIAS).

References

[1] I. Goodfellow, J. Shlens and C. Szegedy, EXPLAINING AND HARNESSING ADVERSARIAL EXAMPLES, 2015.

[2] D. Asimopoulos, P. Radoglou-Grammatikis, I. Makris, V. Mladenov, K. Psannis, S. Goudos and P. Sarigiannidis, "Breaching the defense: Investigating FGSM and CTGAN adversarial attacks on IEC 60870-5-104 AI-enabled Intrusion Detection Systems," *Proceedings of the 18th International Conference on Availability, Reliability and Security,* pp. 1-8, 2023.

[3] W. Brendel, J. Rauber and M. Bethge, DECISION-BASED ADVERSARIAL ATTACKS: RELIABLE ATTACKS AGAINST BLACK-BOX MACHINE LEARNING MODELS, 2018.

[4] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. Celik and A. Swami, Practical Black-Box Attacks against Machine Learning, 2017.

[5] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow and R. Fergus, Intriguing properties of neural networks, 2014.

[6] B. Villarini, P. Radoglou-Grammatikis, T. Lagkas, P. Sarigiannidis and V. Argyriou, "Detection of Physical Adversarial Attacks on Traffic Signs for Autonomous Vehicles," in *2023 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT),* Bali, Indonesia, 2023.

[7] B. Biggio, P. Laskov and B. Nelson, Poisoning Attacks against Support Vector Machines, 2013.

[8] F. Tramèr, F. Zhang, A. Juels, M. Reiter and T. Ristenpart, Stealing Machine Learning Models via Prediction APIs, 2016.

[9] M. Nasr, R. Shokri and A. Houmansadr, Machine Learning with Membership Privacy using Adversarial Regularization, 2018.

[10] B. Thuraisingham, "Trustworthy Artificial Intelligence for Securing Transportation Systems," in *Proceedings of the 29th ACM Symposium on Access Control Models and Technologies*, 2024.

[11] P. Ghosh, "AI-Based Systems for Autonomous Vehicle Traffic Sign Compliance," *Distributed Learning and Broad Applications in Scientific Research,* vol. 9, pp. 84-108, 2023.

[12] M. Lee, J. Yoon and C. Choi, "Adversarial attack vulnerability for multi-biometric authentication system," *Expert Systems,* vol. 41, no. 10, 2024.

[13] K. Nguyen, T. Fernando, C. Fookes and S. Sridharan, "Physical Adversarial Attacks for Surveillance: A Survey," *IEEE Transactions on Neural Networks and Learning Systems,* 2023.

[14] I. Fursov, M. Morozov, N. Kaploukhaya, E. Kovtun, R. Rivera-Castro, G. Gusev, D. Babaev, I. Kireev, A. Zaytsev and E. Burnaev, "Adversarial attacks on deep models for financial transaction records," *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining,* pp. 2868-2878, 2021.

[15] A. Habbal, M. K. Ali and M. Abuzaraida, "Artificial Intelligence Trust, risk and security management (AI trism): Frameworks, applications, challenges and future research directions," *Expert Systems with Applications,* p. 240, 2024.

[16] H. Zhang and K. Ogasawara, "Grad-CAM-based explainable artificial intelligence related to medical text processing," *Bioengineering,* vol. 10, no. 9, p. 1070, 2023.

[17] R. Younisse, A. Ahmad and Q. Abu Al-Haija, "Explaining intrusion detection-based convolutional neural networks using shapley additive explanations (shap)," *Big Data and Cognitive Computing,* vol. 6, no. 4, p. 126, 2022.

[18] A. Javadpour, F. Ja'fari, T. Taleb, M. Shojafar and C. Benzaïd, "A comprehensive survey on cyber deception techniques to improve honeypot performance," *Computers & Security,* 2024.

[19] S. Suhail, M. Iqbal and K. McLaughlin, "Digital Twin-Driven Deception Platform: Vision and Way Forward," *IEEE Internet Computing,* 2024.

[20] H. Wei, H. Tang, X. Jia, Z. Wang, H. Yu, Z. Li, S. Satoh, L. Gool and Z. Wang, "Physical adversarial attack meets computer vision: A decade survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence.*

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

[22] K. Mahima, M. Ayoob and G. Poravi, "Adversarial Attacks and Defense Technologies on Autonomous Vehicles: A Review," *Appl. Comput. Syst.,* vol. 26, no. 2, pp. 96-106, 2021.

[23] C. Szegedy, "Intriguing properties of neural networks," *arXiv,* 2013.

[24] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *arXiv,* 2016.

[25] Y. Li, Z. Liu, J. Zhao, L. Ren, F. Li, J. Luo and B. Luo, "The Adversarial AI-Art: Understanding, Generation, Detection, and Benchmarking," *In European Symposium on Research in Computer Security,* pp. 311-331, 2024.

[26] J. Metzen, T. Genewein, V. Fischer and B. Bischoff, " On detecting adversarial perturbations," *arXiv,* 2017.

[27] Y. Wang, T. Li, S. Li, X. Yuan and W. Ni, "New adversarial image detection based on sentiment analysis," *IEEE Transactions on Neural Networks and Learning Systems,* 2023.

[28] X. Hu, P. Chen and T. Ho, "Radar: Robust ai-text detection via adversarial learning," *Advances in Neural Information Processing Systems,* vol. 36, pp. 15077-15095, 2023.

[29] Y. Diao, N. Zhai, C. Miao, X. Yang and M. Wang, "Vulnerabilities in AI-generated Image Detection: The Challenge of Adversarial Attacks," *arXiv,* 2024.

[30] O. A. Sarcea, "AI & Cybersecurity–connection, impacts, way ahead," *International Conference on Machine Intelligence & Security for Smart Cities (TRUST) Proceedings,* vol. 1, pp. 17-26, 2024.

[31] G. L. Popa, "Risk management, protection, and security of personal data in Romania," *International Conference on Machine Intelligence & Security for Smart Cities (TRUST) Proceedings,* vol. 1, pp. 69-78, 2024.

[32] M. Malatji and A. Tolah, "Artificial intelligence (AI) cybersecurity dimensions: a comprehensive framework for understanding adversarial and offensive AI," *AI and Ethics,* pp. 1-28, 2024.

[33] J. Gu, J. Xiaojun, P. d. Jorge, Y. Wenqain, L. Xinwei, A. Ma, X. Yuan, H. Anjun, K. Ashkan, L. Zhijiang, C. Xiaochun and T. Philip, "A Survey on Transferability of Adversarial Examples across Deep Neural Networks," *arXiv,* 2023.

[34] Y. Wang, T. Sun, S. Li, X. Yuan, W. Ni, E. Hossain and H. Poor, "Adversarial attacks and defenses in machine learning-empowered communication systems and networks: A contemporary survey," *IEEE Communications Surveys & Tutorials,* 2023.

[35] Department of Homeland Security, "Risks and mitigation strategies for adversarial artificial intelligence threats: A DHS S&T study," 2023. [Online]. Available: https://www.dhs.gov/science-and-technology/publication/risks-and-mitigation-strategies-adversarial-artificial-intelligence-threats.

[36] Y. Khaleel, M. Habeeb, A. Albahri, T. Al-Quraishi, O. Albahri and A. Alamoodi, "etwork and cybersecurity applications of defense in adversarial attacks: A state-of-the-art using machine learning and deep learning methods," *Journal of Intelligent Systems,* vol. 33, no. 1, 2024.

[37] G. Boce, "Bias in artificial intelligence," *Smart Cities International Conference (SCIC) Proceedings,* vol. 10, pp. 337-344, 2022.

[38] G. Waizel, "Bridging the AI divide: The evolving arms race between AI-driven cyber attacks and AI-powered cybersecurity defenses," *International Conference on Machine Intelligence & Security for Smart Cities (TRUST) Proceedings,* vol. 1, pp. 141-156, 2024.

[39] M. Nagahisarchoghaei, M. Karimi, S. Rahimi, L. Cummins and G. Ghanbari, "Generative Local Interpretable Model-Agnostic Explanations," *The International FLAIRS Conference Proceedings,* vol. 36, 2023.

[40] Y. Nohara, K. Matsumoto, H. Soejima and N. Nakashima, "Explanation of machine learning models using shapley additive explanation and application for real data in hospital," *Computer Methods and Programs in Biomedicine,* p. 214, 2022.

[41] S. Yin, L. Wang, M. Shafiq, L. Teng, A. Laghari and M. Khan, "G2Grad-CAMRL: an object detection and interpretation model based on gradient-weighted class activation mapping and reinforcement learning in remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* vol. 16, pp. 3583-3598, 2023.

[42] G. Schwalbe and B. Finzel, "A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts," *Data Mining and Knowledge Discovery,* vol. 38, no. 5, pp. 3043-3101, 2024.

[43] N. Ilg, P. Duplys, D. Sisejkovic and M. Menth, "Survey of contemporary open-source honeypots, frameworks, and tools," *Journal of Network and Computer Applications,* 2023.

[44] X. Wang, L. Shi, C. Cao, W. Wu, Z. Zhao, Y. Wang and K. Wang, "Game analysis and decision making optimization of evolutionary dynamic honeypot," *Computers and Electrical Engineering,* 2024.

[45] S. Mihai, M. Yaqoob, D. Hung, W. Davis, P. Towakel, M. Raza and H. Nguyen, "Digital twins: A survey on enabling technologies, challenges, trends and future prospects," *IEEE Communications Surveys & Tutorials,* vol. 24, no. 4, pp. 2255-2291, 2022.

[46] I. Sarker, H. Janicke, A. Mohsin, A. Gill and L. Maglaras, "Explainable AI for cybersecurity automation, intelligence and trustworthiness in digital twin: Methods, taxonomy, challenges and prospects," *ICT Express,* 2024.

[47] R. Dwivedi, D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel and R. Ranjan, "Explainable AI (XAI): Core ideas, techniques, and solutions," *ACM Computing Surveys,* vol. 55, no. 9, pp. 1-33, 2023.

[48] S. Wang and Y. Zhang, "Grad-CAM: understanding AI models," *Comput. Mater. Contin,* vol. 76, no. 2, pp. 1321-1324, 2023.

[49] V. Pantelakis, P. Bountakas, A. Farao and C. Xenakis, "Adversarial machine learning attacks on multiclass classification of iot network traffic," *Proceedings of the 18th International Conference on Availability, Reliability and Security,* pp. 1-8, 2023.

[50] A. Farao, C. Ntantogian, S. Karagiannis, E. Magkos, A. Dritsa and C. Xenakis, "NITRO: an Interconnected 5G-IoT Cyber Range," *Proceedings of the 19th International Conference on Availability, Reliability and Security,* pp. 1-6, 2024.

[51] G. Petihakis, A. Farao, P. Bountakas, A. Sabazioti, J. Polley and C. Xenakis, "AIAS: AI-ASsisted cybersecurity platform to defend against adversarial AI attacks," *Proceedings of the 19th International Conference on Availability, Reliability and Security,* pp. 1-7, 2024.

[52] M. Karatisoglou, A. Farao, V. Bolgouras and C. Xenakis, "BRIDGE: BRIDGing the gap bEtween CTI production and consumption," *2022 14th International Conference on Communications (COMM),* pp. pp. 1-6, 2022.

[53] A. Muñoz, A. Farao, J. Correia and C. Xenakis, "P2ISE: preserving project integrity in CI/CD based on secure elements," *Information,* vol. 12, no. 9, p. 357, 2021.

[54] A. Muñoz, A. Farao, J. Correia and C. Xenakis, "ICITPM: integrity validation of software in iterative continuous integration through the use of Trusted Platform Module (TPM)," in *Computer Security: ESORICS 2020 International Workshops, DETIPS, DeSECSys, MPS, and SPOSE,* Guildford, UK, 2020.

[55] G. Suciu, A. Farao, G. Bernardinetti, I. Palamà, M. Sachian, A. Vulpe and C. Xenakis, "SAMGRID: security authorization and monitoring module based on SealedGRID platform," *Sensors,* vol. 22, no. 17, 2022.

[56] T. J. Holt, "On the value of honeypots to produce policy recommendations," *Criminology & Pub,* 2017.

[57] C. Kelly, N. Pitropakis, A. Mylonas, S. McKeown and W. Buchanan, "A Comparative Analysis of Honeypots on Different Cloud Platforms," *Sensors,* vol. 21, no. 7, 2021.

[58] G. Waizel, "Using a modern honeypot model to defend smart cities and provide early detection to APT and ransomware attacks," *Smart Cities International Conference (SCIC) Proceedings,* vol. 10, pp. 33-46.

[59] T. Holz and F. Raynal, "Detecting honeypots and other suspicious environments," *Proceedings from the sixth annual IEEE SMC information assurance workshop,* 2005.

[60] I. Mokube and M. Adams, "Honeypots: concepts, approaches, and challenges," *Proceedings of the 45th annual southeast regional conference,* 2007.

[61] M. L. Bringer, C. A. Chelmecki and H. Fujinoki, "A survey: Recent advances and future trends in honeypot research," *International Journal of Computer Network and Information Security,* vol. 4, no. 10, 2012.

[62] C. Kreibich and J. Crowcroft, "Honeycomb: creating intrusion detection signatures using honeypots," *ACM SIGCOMM computer communication review,* vol. 34, no. 1, pp. 51-56, 2004.

[63] F. Wenjun, "Enabling an anatomic view to investigate honeypot systems: A survey," *IEEE Systems Journal,* vol. 12, no. 4, pp. 3906-3919, 2017.

[64] S. Almotairi, "A technique for detecting new attacks in low-interaction honeypot traffic," *2009 Fourth International Conference on Internet Monitoring and Protection. IEEE,* 2009.

[65] A. Sharma, "Honeypots in Network Security," *Int. J. Technol. Res. Appl,* vol. 7, no. 12, 2013.

[66] I. Koniaris, G. Papadimitriou and P. Nicopolitidis, "Analysis and visualization of SSH attacks using honeypots," *Eurocon 2013,* 2013.

[67] M. Anirudh, T. Arul and J. N. Daniel, "Use of honeypots for mitigating DoS attacks targeted on IoT networks," *2017 International conference on computer, communication and signal processing (ICCCSP),* 2017.

[68] V. Sethia and A. Jeyasekar, "Malware capturing and analysis using dionaea honeypot," *2019 International Carnahan Conference on Security Technology (ICCST),* 2019.

[69] Jiao Ma, Kun Chai, Yao Xiao, Tian Lan and Wei Huang, "High-Interaction Honeypot System for SQL Injection Analysis," *2011 International Conference of Information Technology, Computer Engineering and Management Sciences,* vol. 3, 2011.

[70] J. Franco, A. Aris, B. Canberk and A. S. Uluagac, "A survey of honeypots and honeynets for internet of things, industrial internet of things, and cyber-physical systems," *IEEE Communications Surveys & Tutorials,* vol. 23, no. 4, pp. 2351-2383, 2021.

[71] M. Zemene and P. Avadhani, "Implementing high interaction honeypot to study SSH attacks," *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI),* 2015.

[72] T. Arshad and M. Santhosh, "AI-Enabled Honeypot." Journal of Network & Information Security," vol. 11, no. 2, 2023.

[73] C. Guan, H. Liu, G. Cao, S. Zhu and T. La Porta, "HoneyIoT: Adaptive High-Interaction Honeypot for IoT Devices Through Reinforcement Learning," *Proceedings of the 16th ACM Conference on Security and Privacy in Wireless and Mobile Networks,* 2023.

[74] E. Gizzarelli, "Honeypot and Generative AI.," *Diss. Politecnico di Torino,* 2024.

[75] "cowrie," [Online]. Available: https://github.com/cowrie/cowrie .

[76] "Kippo," [Online]. Available: https://github.com/desaster/kippo.

[77] "dionaea," [Online]. Available: https://github.com/DinoTools/dionaea.

[78] A. Kostopoulos, I. Chochliouros, T. Apostolopoulos, C. Patsakis, G. Tsatsanifos, M. Anastasiadis and B. Tran, "Realising honeypot-as-a-service for smart home solutions," *2020 5th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM),* pp. 1-6, 2020.

[79] "tpotce," [Online]. Available: https://github.com/telekom-security/tpotce.

[80] K. Wang, M. Du, S. Maharjan and Y. Sun, "Strategic honeypot game model for distributed denial of service attacks in the smart grid," *IEEE Transactions on Smart Grid,* vol. 8, no. 5, pp. 2474-2482, 2017.

[81] B. Farinholt, M. Rezaeirad, P. Pearce, H. Dharmdasani, H. Yin, S. Le Blond and K. Levchenko, "To Catch a Ratter: Monitoring the Behavior of Amateur DarkComet RAT Operators in the Wild," *Security and Privacy (SP),* pp. 770-787, 2017.

[82] A. Pokhrel, K. Vikash and R. Colomo-Palacios, "Digital twin for cybersecurity incident prediction: A multivocal literature review," *Proceedings of the IEEE/ACM 42nd international conference on software engineering workshops,* 2020.

[83] J. Lopez, J. Rubio and C. Alcaraz, "Digital Twins for Intelligent Authorization in the B5G-Enabled Smart Grid," *IEEE Wireless Communications,* vol. 28, no. 2, pp. 48-55, 2021.

[84] R. Faleiro, L. Pan, S. Pokhrel and R. Doss, "Digital twin for cybersecurity: Towards enhancing cyber resilience," *Broadband Communications, Networks, and Systems: 12th EAI International Conference,* pp. 57-76, 2022.

[85] "Platform Stack Architectural Framework: An Introductory Guide," [Online]. Available: https://www.digitaltwinconsortium.org/platform-stack-architectural-fram-formework-an-introductory-guide-form/.