AI in Cybersecurity: Activities of the CINI-AIIS Lab at University of Naples Federico II

Antonino Ferraro[1], Antonio Galli[1,*], Valerio La Gatta[1,2], Lidia Marassi[1], Stefano Marrone[1], Vincenzo Moscato[1], Marco Postiglione[1,2], Carlo Sansone[1] and Giancarlo Sperli[1]

[1]University of Naples Federico II, Via Claudio 21, Naples, 80125, Italy
[2]Northwestern University, Department of Computer Science, McCormick School of Engineering and Applied Science, 2233 Tech Dr, Evanston, IL 60208, United States

Abstract

Artificial intelligence (AI) is revolutionizing various industries, including cybersecurity, by emulating human intelligence to address complex threats. In the cybersecurity domain, AI offers significant potential, bolstering defense mechanisms, optimizing threat detection, and advancing incident response capabilities. AI-powered systems can analyze vast datasets to identify anomalies, predict cyberattacks, and enhance overall security posture. Machine Learning (ML), a subset of AI, enables systems to learn from data and make informed decisions, such as predicting optimal security measures based on threat intelligence and operational context. Deep Learning (DL), another ML subset, harnesses Artificial Neural Networks (ANNs) to process intricate data patterns and provide accurate threat assessments. DL, especially through Convolutional Neural Networks (CNNs), is transforming cybersecurity by extracting meaningful features from network traffic and log data for anomaly detection and threat hunting. Moreover, DL integrated with Natural Language Processing (NLP) streamlines tasks like threat intelligence analysis and incident response coordination. The versatility of AI underscores its pivotal role in cybersecurity, driving resilience enhancements and fostering proactive defense strategies. In this paper, we highlight AI projects in the cybersecurity sector from the University of Naples Federico II node of the CINI-AIIS Lab, showcasing their innovative contributions to cyber defense.

Keywords

Artificial Intelligence, Cybersecurity, Deep Learning, Machine Learning

1. Introduction

Artificial intelligence (AI) is a transformative force across various industries, providing a paradigm shift in cybersecurity practices. Within the cybersecurity domain, AI is heralding significant advancements, redefining defensive strategies, amplifying threat detection capabilities, and refining incident response mechanisms. By harnessing AI technologies, organizations can fortify their defensive postures, anticipate and mitigate cyber threats proactively, and elevate overall security resilience.

At the core of AI's impact on cybersecurity lies its capacity to analyze vast and diverse datasets, enabling the identification of anomalies, prediction of emerging threats, and optimization of security measures. Machine Learning (ML), a pivotal subset of AI, equips systems with the ability to learn from data, thereby enhancing decision-making processes based on evolving threat landscapes and operational contexts. Deep Learning (DL), another cornerstone of AI, leverages sophisticated Artificial Neural Networks (ANNs) to discern intricate patterns within data, furnishing precise threat assessments and actionable insights. Particularly through Convolutional Neural Networks (CNNs), DL revolutionizes cybersecurity by extracting salient features from network traffic and log data, facilitating anomaly detection, threat prediction, and forensic analysis.

Moreover, the fusion of DL with Natural Language Processing (NLP) streamlines critical cybersecurity tasks, such as threat intelligence analysis, malware detection, and incident response coordination. By comprehensively analyzing textual data, NLP-powered systems augment analysts' capabilities, enabling rapid threat identification and proactive response measures.

The adaptable and multifaceted nature of AI positions it as a cornerstone of cybersecurity, driving innovation, resilience, and agility in the face of evolving threats. In this paper, we present a comprehensive overview of AI initiatives in cybersecurity, drawing from projects conducted at the University of Naples Federico II node of the CINI-AIIS Lab. Through these endeavors, we showcase the transformative potential of AI in bolstering cyber defense strategies and safeguarding digital ecosystems against emerging threats.

2. Interpreting AI Models for Behavioral Malware Detection

In the past decade, the landscape of cyber threats to Information Systems has undergone a remarkable trans-

Ital-IA 2024: 4th National Conference on Artificial Intelligence, organized by CINI, May 29-30, 2024, Naples, Italy
*Corresponding author.
antonio.galli@unina.it (A. Galli)

formation, driven largely by the widespread adoption of Internet of Things (IoT) devices and Cloud Computing technologies. This proliferation has provided cybercriminals with a fertile ground for launching a multitude of attacks, ranging from the insertion of unwanted advertisements into websites to the clandestine exfiltration of sensitive data for illicit financial gains. At the forefront of these attacks are various forms of malicious software, collectively referred to as malware, which pose significant challenges to the security and integrity of digital systems. Examples of such malware include trojans, backdoors, spyware, and worms, each designed with the explicit purpose of exploiting vulnerabilities in target systems ([1]).

The detection of malware represents a formidable research endeavor, compounded by the ever-evolving sophistication of cyber threats. As Cyber Security (CS) researchers develop new detection techniques, malware authors respond in kind, continually refining their strategies to evade detection ([2, 3]). In this perpetual arms race, traditional antivirus software programs, reliant on signature-based detection mechanisms, have struggled to keep pace with the rapidly evolving threat landscape. Signature-based detection relies on identifying known patterns or signatures of malicious code within a database, often leading to a cat-and-mouse game where malware authors employ advanced evasion techniques such as code obfuscation to circumvent detection ([4, 5]).

To address the shortcomings of signature-based detection, researchers have explored alternative approaches that focus on analyzing malware behavior, rather than static code signatures. These approaches can be broadly categorized into Static Malware Detection (SMD) and Behavioral Malware Detection (BMD). SMD techniques analyze the static characteristics of malware, such as its byte-code structure, while BMD approaches monitor the dynamic behavior of malware at runtime, particularly the sequence of Application Programming Interface (API) calls made by the software to the underlying operating system ([6]). This behavioral analysis provides valuable insights into the actions performed by malware, offering a more comprehensive understanding of its capabilities and intentions.

However, the complexity and variability of modern malware present significant challenges to both SMD and BMD approaches. Static analysis techniques are vulnerable to evasion tactics such as dynamic code linking and encryption, while behavioral analysis can be computationally intensive and time-consuming ([7, 8]). In response to these challenges, researchers have turned to advanced Machine Learning (ML) and Deep Learning (DL) techniques to enhance the effectiveness of malware detection systems ([9, 10, 7]). These approaches leverage the power of neural networks to automatically learn complex patterns and features from raw data, offering promising avenues for improving detection accuracy and efficiency.

Despite their impressive performance, ML and DL-based detection systems often lack transparency and interpretability, raising concerns about their trustworthiness and reliability in real-world applications. To address these concerns, researchers have begun exploring the field of eXplainable Artificial Intelligence (XAI), which focuses on developing models and techniques that can provide human-understandable explanations for AI-driven decisions ([11]). In the context of malware detection, XAI methodologies aim to elucidate the underlying reasoning behind classification decisions, offering valuable insights into the features and patterns driving the detection process.

While XAI approaches have shown promise in enhancing the explainability of malware detection systems, their application to Behavioral Malware Detection (BMD) remains relatively unexplored, particularly in the context of deep sequential neural networks. This gap in research underscores the need for comprehensive investigations into the explainability of BMD systems, especially as they become increasingly reliant on advanced DL techniques. In our research, we present a novel XAI framework for BMD, leveraging a range of state-of-the-art techniques to provide transparent and interpretable explanations for classification decisions. Through extensive experimentation on publicly available datasets, we evaluate the effectiveness and robustness of our framework, shedding light on its utility and potential limitations in real-world cybersecurity applications.

More in details, our methodology builds upon a pipeline composed by three steps: the *sequence preprocessing* module aims to standardize the data format, the *model* is a classification learner that exploits the sequence structure of input data to perform the classification and *the explainer* generates the explanation supporting the model's prediction. Our methodological workflow is summarized in Fig. 1.

To sum up, we introduced an Explainable Artificial Intelligence (XAI) framework for behavioral malware detection. We aimed to assess the effectiveness of four XAI methods within a sequence-based deep learning model and their relevance in contemporary cybersecurity applications.

Our experiments demonstrated the feasibility of various XAI techniques in explaining the decisions of LSTM-based classifiers, considering both explanation quality and computational efficiency. While our focus was on local explanations for individual samples, global explanations were not addressed.

However, limitations exist, particularly regarding the lack of qualitative metrics to directly evaluate XAI effectiveness and the potential influence of domain-specific factors on our findings.

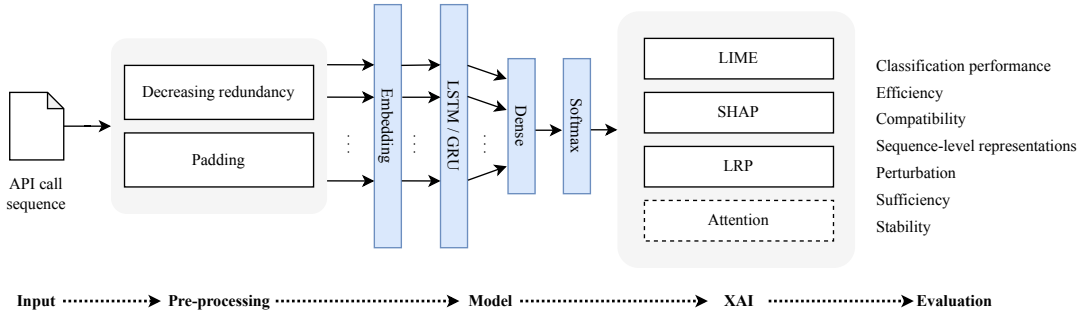

Figure 1: Methodological workflow. The pre-processing step aims to standardize the data format. The model classifies the input sequence as malware/goodware, and the explainer generates the explanation. The models are then evaluated in terms of classification performance, efficiency and explanations quality.

Future research will explore additional XAI methods and assess the robustness of our framework against adversarial attacks. We also plan to investigate whether explanations can enhance classification performance and assist in identifying systematic errors in predictive models. Real-world scenarios will be considered to evaluate the practical utility of explanations in aiding expert analysts.

3. Autoencoder-Based Deep Learning Pipeline for Network Anomaly Detection

In recent years, the rapid expansion of interconnected devices, like those found in IoT and Cloud networks, has highlighted the urgent need for strong network security assessments. One crucial aspect of addressing this challenge is detecting network anomalies, which serve as important indicators of network intrusions, privacy breaches, system damage, and fraudulent activities. Deep neural networks, known for their ability to learn intricate anomaly patterns from data, have become increasingly popular in this field. However, their effectiveness can be hampered by the unique characteristics of network traffic data, which is sparse, noisy, and often imbalanced due to the multitude of devices and internet applications generating it. Anomalies typically occur in only a small fraction of instances, ranging from 0.001% to 1%. In our research, we tackle these challenges with a focused approach. Initially, we use an autoencoder (AE) to identify instances of anomalous behavior. Then, these anomalies are classified by an attack classifier based on their specific type. We have tested our framework on a large-scale dataset consisting of real-world network traffic data, yielding promising results.

Our proposed framework, as depicted in Figure 2, operates at a high level by processing session description attributes s_i (such as port number and bytes transferred) and determining whether the input is benign or represents an attack. In cases of an attack, the output y_i identifies the specific type of attack (e.g., DDoS, sweep).

Denoising Autoencoder (DAE): The DAE module processes the i-th session $s_i \in \mathbb{R}^n$ and outputs its latent representation $\tilde{x}_i \in \mathbb{R}^k$ and the reconstructed instance $\tilde{s}_i \in \mathbb{R}^n$. The latent representation can be considered as the DAE features, while the reconstructed instance represents how the input session might be generated from the latent space.

Reconstruction Error (RE) Module: The RE module utilizes the output of the DAE, $\tilde{s}_i$, to calculate the reconstruction error $e_i \in \mathbb{R}$. This error is indicative of the autoencoder's proficiency in interpreting the input session - a higher error suggests a poorer representation. The RE module assesses the similarity between s_i and $\tilde{s}_i$ using various metrics $m()$, such as cosine similarity or dot product, with empirical evidence favoring the former for enhanced results.

Threshold Module (TRH): The TRH module concatenates the reconstruction error e_i with the latent representation $\tilde{x}_i$, forming a comprehensive feature vector for the input instance. It functions as a binary classifier within a multilayer perceptron architecture, discerning if the DAE has recognized s_i as akin to the benign instances it was trained on:

$$f : \tilde{x}_i \in \mathbb{R}^k \rightarrow \{0, 1\} \tag{1}$$

Here, a positive class indicates a benign session, while a negative class signals an attack, the specifics of which are determined by the AC module.

Attack Classifier (AC): In tandem with the TRH computation, the AC module also receives the concatenated vector of e_i and $\tilde{x}_i$. The AC module employs a multi-class tabular classifier (such as a random forest or support vector machine) that can be trained using standard supervised machine learning methods. It assigns the attack

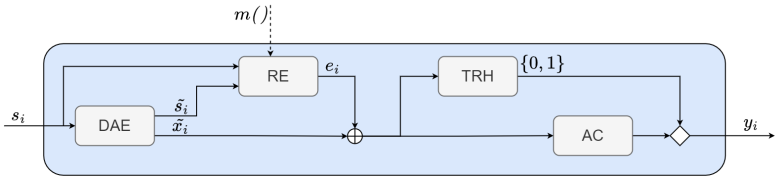

Figure 2: Overview of proposed NAD pipeline.

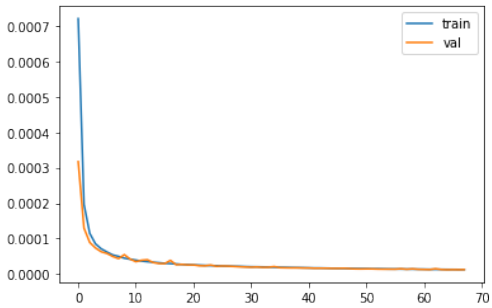

Figure 3: DAE reconstruction error on training and validation splits. On the x axis we report the increasing number of epochs, while MSE values are reported on the y axis.

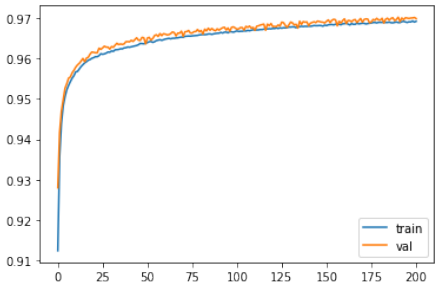

Figure 4: TRH accuracy on training and validation splits. On the x axis we report the increasing number of epochs, while accuracy values are reported on the y axis.

typology to the input instance, with the choice of classification algorithm impacting overall performance, as detailed in the experimental section. The final decision of the framework is derived by considering the outputs of both the TRH and AC modules. If the TRH output is zero, indicating successful reconstruction by the DAE, the input instance is classified as benign. If not, the input instance is classified according to the attack type predicted by the AC module. This approach leverages the DAE's ability to recognize benign sessions, a capability honed through extensive training on numerous instances, while the AC module provides the specificity in attack typology classification when an attack is presumed.

Our dataset has been provided with the NAD2021 challenge [12], where participants are provided with traffic records from three specific dates, classified as either normal traffic or a specific type of network attack. The challenge focuses on two primary types of attacks: (1) *probing attacks*, that involve attempts to extract data from a targeted network, and (2) DDOS-Smurf attacks, which are characterized by the use of numerous ICMP flows, aimed at overwhelming and halting traffic to a specific destination IP address.

The DAE module was trained using an early stopping mechanism, halting after three epochs without MSE improvement on the validation set. Figure 3 show that training stops at 69 epochs and the model easily learns to reconstruct input samples. The final MSE scores were 1.2944e-5 for training and 1.2402e-5 for validation. Additionally, further training for five epochs using both training and validation data reduced the training MSE to 1.1759e-5.

Table 1
Attacks Classifier, validation performance

Anomaly	Precision	Recall	F1
DDoS	0.99	1.00	0.99
IP sweep	1.00	1.00	1.00
Nmap sweep	0.98	0.87	0.92
Port sweep	0.99	0.99	0.99

The TRH model, integrating latent features from the DAE and its reconstruction error, was trained to classify samples as Normal (0) or Anomalous (1), using a similar early stopping strategy set at 10 epochs. Figure 4 show that training stops at epoch 202 with a training accuracy $Acc_{train} = 0.9697$ and validation accuracy $Acc_{val} = 0.9698$. These results indicate the model's proficiency in differentiating between anomalous and normal samples.

The AC module, tasked with classifying attack samples identified by the TRH, was trained using a RandomForest classifier. Performance metrics, including Precision, Recall, and F1 scores, are detailed in the classification report. The confusion matrix provides further insights into the classifier's performance across different attack types. We report results in Table 1 (Precision, Recall and F1 scores) and Table 2 (confusion matrix).

Table 2
Attacks classifier, validation confusion matrix

	DDoS	IP sweep	Nmap sweep	Port sweep
DDoS	374	1	0	0
IP sweep	2	38310	0	172
Nmap sweep	1	4	116	12
Port sweep	2	109	2	12253

Table 3
Test performance of DAE+TRH modules distinguishing anomalous and normal samples

Class	Precision	Recall	F1
Normal	1.00	0.96	0.98
Anomaly	0.47	0.98	0.63

Table 4
Test performance of the entire DAE+TRH+AC pipeline

Class	Precision	Recall	F1
DDoS	0.11	0.52	0.19
Normal	1.00	0.96	0.98
IP sweep	0.53	0.99	0.69
Nmap sweep	0.96	0.83	0.89
Port sweep	0.34	0.95	0.50

The final test assessed the combined performance of the DAE, TRH, and AC modules on the test set. Given the unbalanced nature of the data, Precision and Recall were key metrics for evaluating the DAE+TRH's ability to distinguish between normal and anomalous samples. While these modules demonstrated high quality in differentiating negatives from positives, there were limitations in identifying all anomalies. The cumulative errors from the DAE+TRH and AC modules are reflected in the overall system performance. The aggregated $F_{\alpha\beta}$ score, evaluating the system across all classes, was recorded as 0.577, indicating areas for improvement in the pipeline's ability to accurately classify various types of network activities.

In conclusion, we introduced a streamlined and effective framework for Network Anomaly Detection (NAD). Our approach involves two main phases: (1) identifying anomalies using latent features generated by a Deep Denoising Autoencoder, and (2) classifying these anomalies with a multi-label classifier. Despite potential error propagation within the pipeline, our approach has shown promising results. However, we observed a limitation in the performance of the Threshold module (TRH), particularly in detecting attack samples, due to dataset imbalance. Future research will focus on implementing class-balancing techniques to improve the TRH module's effectiveness and enhance the overall system performance.

4. AI Act and Biometrics

As AI becomes more integrated into daily life, cybersecurity emerges as a critical concern. The AI Act, the first global law on AI usage, serves as a key regulatory framework within the European Union, emphasizing ethical considerations in cybersecurity. This law seeks a balance between technological innovation and the protection of core ethical values, ensuring AI is used responsibly. Particularly important within the AI Act is the role of cybersecurity for high-risk AI systems, which requires a comprehensive security approach. One significant challenge addressed by the AI Act is the management of biometrics, acknowledging their sensitive nature and the privacy and security implications for individuals. The act is particularly concerned with the ethical use of biometric data, such as fingerprints, and facial and vocal recognition, due to the personal data protection it necessitates. To regulate the deployment of facial and biometric recognition technologies in public spaces, the AI Act sets strict rules, allowing exceptions only in well-defined scenarios like locating missing persons or preventing serious crimes [13].

While the AI Act represents a significant step forward in balancing the benefits of artificial intelligence with the protection of fundamental rights, it also makes even more complex the landscape of challenges that remain. Indeed, on one hand, stringent regulations are essential for managing the risks associated with AI technologies and ensuring they adhere to ethical standards. On the other hand, continuous research in the field of AI and biometrics is critical. The need for advancing research in biometrics is recognized globally, to the extent that numerous international competitions have been established to challenge researchers in identifying fake biometrics. Over the years, the Naples' CINI AI-IS node has made significant contributions to the field of fake fingerprint detection. It has actively participated in several editions of LIVDET[1], an international competition that challenges researchers with the task of distinguishing between live and fake fingerprints created through diverse techniques and spoofing materials. Our team has achieved notable success in the last two editions, securing first place in one and second place in another. These accomplishments were made possible through our innovative use of adversarial learning techniques, which allowed us to perform a synthetic data augmentation able to improve the overall performance of a liveness detector [14] achieving an accuracy over 90% on two dataset. More recently, exploiting the experience matured over the years, we also developed a new fake fingerprint crafting strategy that can be used to physically cast a fake fingerprint able to bypass AI-based liveness detectors [15].

[1]https://sites.unica.it/livdet/

These results not only anticipate future cybersecurity threats but also aid in formulating effective defence mechanisms. To address this need while also protecting people from unwanted misuses, we advocate that one of the major challenges in the field of AI is education, to promote a deeper understanding of the risks and ethical implications of AI and enable people to participate in an informed and conscious manner in public debate and decision-making regarding the use and regulation of these technologies. In pursuing a balance between technological innovation and the protection of fundamental rights, it seems necessary to promote an open and inclusive dialogue involving both developers and civil society stakeholders [16].

Acknowledgments

This work was supported in part by the Piano Nazionale Ripresa Resilienza (PNRR) Ministero dell'Università e della Ricerca (MUR) Project under Grant PE0000013-FAIR

References

[1] S. Yan, J. Ren, W. Wang, L. Sun, W. Zhang, Q. Yu, A survey of adversarial attack and defense methods for malware classification in cyber security, IEEE Communications Surveys & Tutorials 25 (2023) 467–496. doi:10.1109/COMST.2022.3225137.

[2] N. Galloro, M. Polino, M. Carminati, A. Continella, S. Zanero, A Systematical and longitudinal study of evasive behaviors in windows malware, Computers & Security 113 (2022) 102550. doi:https://doi.org/10.1016/j.cose.2021.102550.

[3] F. Zhong, X. Cheng, D. Yu, B. Gong, S. Song, J. Yu, MalFox: Camouflaged Adversarial Malware Example Generation Based on Conv-GANs Against Black-Box Detectors, IEEE Transactions on Computers (2023) 1–14. doi:10.1109/TC.2023.3236901.

[4] Z. Bazrafshan, H. Hashemi, S. M. H. Fard, A. Hamzeh, A survey on heuristic malware detection techniques, in: The 5th Conference on Information and Knowledge Technology, IEEE, 2013, pp. 113–120. doi:10.1109/IKT.2013.6620049.

[5] B. Cheng, J. Ming, E. A. Leal, H. Zhang, J. Fu, G. Peng, J.-Y. Marion, {Obfuscation-Resilient} executable payload extraction from packed malware, in: 30th USENIX Security Symposium (USENIX Security 21), 2021, pp. 3451–3468.

[6] M. Alazab, R. Layton, S. Venkatraman, P. Watters, Malware detection based on structural and behavioural features of api calls, in: International cyber resilience conference (1st: 2010), Edith Cowan University, 2010, pp. 1–10.

[7] M. G. Gaber, M. Ahmed, H. Janicke, Malware detection with artificial intelligence: A systematic literature review, ACM Computing Surveys (2023). doi:10.1145/3638552.

[8] A. Damodaran, F. Di Troia, C. A. Visaggio, T. H. Austin, M. Stamp, A comparison of static, dynamic, and hybrid analysis for malware detection, Journal of Computer Virology and Hacking Techniques 13 (2017) 1–12. doi:https://doi.org/10.1007/s11416-015-0261-z.

[9] F. O. Catak, A. F. Yazı, O. Elezaj, J. Ahmed, Deep learning based sequential model for malware analysis using windows exe api calls, PeerJ Computer Science 6 (2020) e285. URL: https://doi.org/10.7717/peerj-cs.285. doi:10.7717/peerj-cs.285.

[10] G. M., S. C. Sethuraman, A comprehensive survey on deep learning based malware detection techniques, Computer Science Review 47 (2023) 100529. doi:https://doi.org/10.1016/j.cosrev.2022.100529.

[11] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, F. Herrera, Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence, Information Fusion 99 (2023) 101805. doi:https://doi.org/10.1016/j.inffus.2023.101805.

[12] L. Chen, S.-E. Weng, C.-J. Peng, H.-H. Shuai, W.-H. Cheng, Zyell-nctu nettraffic-1.0: A large-scale dataset for real-world network anomaly detection, 2021. URL: https://arxiv.org/abs/2103.05767. doi:10.48550/ARXIV.2103.05767.

[13] T. Madiega, Artificial intelligence act, European Parliament: European Parliamentary Research Service (2021).

[14] A. Galli, M. Gravina, S. Marrone, D. Mattiello, C. Sansone, Adversarial liveness detector: Leveraging adversarial perturbations in fingerprint liveness detection, IET Biometrics 12 (2023) 102–111.

[15] R. Casula, G. Orrù, S. Marrone, U. Gagliardini, G. L. Marcialis, C. Sansone, Realistic fingerprint presentation attacks based on an adversarial approach, IEEE Transactions on Information Forensics and Security (2023).

[16] J. Borenstein, A. Howard, Emerging challenges in ai and the need for ai ethics education, AI and Ethics 1 (2021) 61–65.