AI-driven solutions for cybersecurity: comparative analysis and ethical aspects

Andreea DINU, Paul Cristian VASILE, Alexandru GEORGESCU

National Institute for Research and Development in Informatics – ICI Bucharest, Romania

andreea.dinu@ici.ro, paul.vasile@ici.ro, alexandru.georgescu@ici.ro

Abstract: In an era where artificial intelligence (AI) is rapidly shaping the future of technology, it is becoming increasingly integrated and utilized across various fields, unlocking a wide range of applications and use cases. Among these, the field of cybersecurity will benefit significantly, as AI has the potential to enhance traditional methods of detection and prevention. This paper aims to present the latest advances in AI technology, exploring how AI can be synergistically connected with cybersecurity and highlighting the types of applications already developed in this area. Additionally, this paper will discuss the ethical considerations and current challenges associated with the use of AI to ensure a more secure environment.

Keywords: AI Techniques, Cybersecurity, Large Language Models (LLMs), Transformers.

Soluții bazate pe Inteligența Artificială (IA) pentru securitatea cibernetică: analiză comparativă și aspecte etice

Rezumat: Într-o eră în care inteligența artificială (IA) modelează rapid viitorul tehnologiei, aceasta devine tot mai integrată și utilizată în diferite domenii, deblocând astfel o gamă largă de aplicații și cazuri de utilizare. Dintre acestea, domeniul securității cibernetice va beneficia semnificativ, deoarece inteligența artificială are potențialul de a îmbunătăți metodele tradiționale de detectare și prevenire. Această lucrare își propune să prezinte cele mai recente progrese în tehnologia IA, explorând modul în care AI poate fi conectată sinergic cu securitatea cibernetică și evidențiind tipurile de aplicații deja dezvoltate în acest domeniu. În plus, această lucrare va discuta considerațiile etice și provocările actuale asociate cu utilizarea inteligenței artificiale pentru a asigura un mediu cât mai sigur.

Cuvinte cheie: inteligență artificială, securitate cibernetică, modele de limbaj mari, transformatori.

1. Introduction

The swift expansion of digital technology has resulted in a surge of data and communication, fundamentally altering the ways that individuals and businesses function. Unprecedented benefits have come from this transition, but it has also shown weaknesses that bad actors could take advantage of. Because of the rising sophistication of cyber-attacks, novel approaches to cybersecurity are required. In this fight, artificial intelligence (AI) has shown to be a potent ally, able to adapt and learn from enormous datasets to identify and neutralize threats instantly (Aslam, 2024).

There has never been a greater demand for flexible, reliable and efficient cyber security solutions. In this environment, it has become evident that the introduction and integration of Artificial Intelligence (AI) in cyber security has the potential to be transformative. The ability of artificial intelligence to replicate and, possibly, even exceed human cognitive processes, has made it obvious that AI is an essential tool for enhancing cyber security. AI can predict with previously unheard-of accuracy, can react to new knowledge, and can extract patterns from enormous databases using sophisticated techniques (Kumar et al., 2023).

In all facets of the digital era, artificial intelligence (AI) has become one of the most important technologies. Technology development in cybersecurity is advancing quickly to use AI to solve security-related concerns. AI-enabled security systems are more versatile, strong, and flexible than traditional cybersecurity solutions. Experts believe that AI security systems will assist to improve cybersecurity performance and defense and ultimately have a big impact on the

https://doi.org/10.33436/v34i3y202403

cybersecurity environment, even though AI technology is still underdeveloped and its application in cybersecurity is still in its infancy. Meanwhile, the attack element is also impacted by the use of AI in cybersecurity. The primary objective of cyberattacks will remain the same, even if artificial intelligence (AI) has a greater influence in cybersecurity. AI-powered attacks not only take down systems and steal data but may also manipulate data to change people's behavior. Secondly, cyberattacks driven by AI won't be employed everywhere. In comparison to conventional cyberattacks, AI attacks require significant resources such as time and money. As a result, rather than focusing on individuals, large-scale incidents using AI technology will use far more sophisticated strategies to target businesses, government organizations and individuals. Thirdly, although experts cannot agree on a specific timeline, they do believe that AI-powered cyberattacks will happen soon (Kim & Park, 2024).

Artificial intelligence (AI) continues to develop at an accelerated rate, and given the current trends, cybersecurity experts and developers need to be aware of potential risks as well as how they can apply methods based on AI to mitigate and prevent cyberattacks. This study aims to explore the potential of integrating artificial intelligence (AI) into cybersecurity, presenting several already developed and placed into practice solutions. The paper will additionally go through the benefits and limitations of adopting AI, particularly as it involves cybersecurity solutions.

2. Related work

Integrating artificial intelligence (AI) into the cybersecurity field is of paramount importance due to the increasingly sophisticated nature of cyber threats. Traditional cybersecurity methods, while foundational, are often inadequate to counter the rapidly evolving tactics employed by malicious actors. AI brings a transformative capability to the domain of cybersecurity through many pivotal mechanisms. The importance of developing numerous AI applications in cybersecurity cannot be overstated. Each application addresses specific aspects of the cybersecurity landscape, collectively enhancing the ability to protect against diverse and sophisticated threats. As cyber threats continue to evolve, the innovation and advancement of AI technologies will be crucial for maintaining robust defense mechanisms. The following sections will present algorithms and applications proposed by experts for the enhancement of the cyber domain, as well as mechanisms for defense and countermeasures.

In the paper by (Aslan & Yilmaz, 2021), authors proposed a deep learning-based malware classification framework. For malware categorization, this system offers a mixed deep neural network architecture. The presented method was testet on Malimg, Microsoft BIG 2015, and Malevis datasets. The suggested system's technique, as seen in Figure 1, consists of three primary components. Firstly, a number of comprehensive datasets are used in the process of gathering malware data. Second, pre-trained networks are used to extract the features of both low- and high-level malware. Ultimately, the deep neural networks architecture's training phase is carried out using a supervised learning technique.

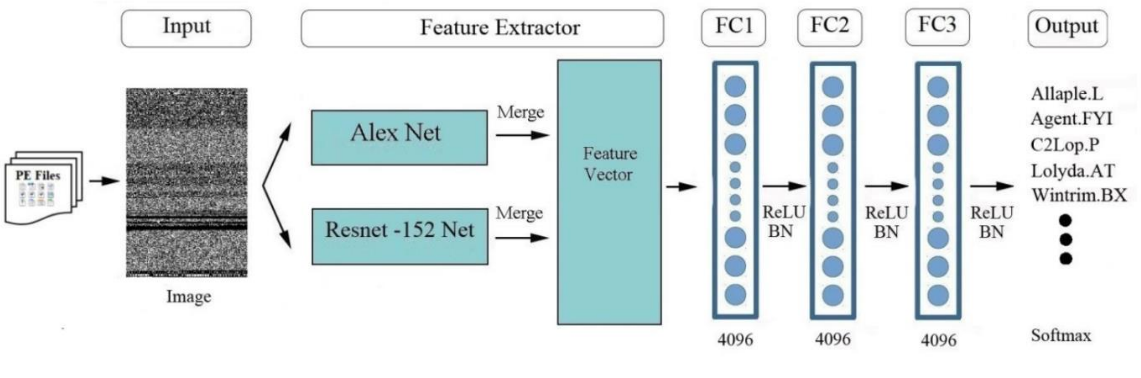

Figure 1. Methodology for malware classification (Aslan & Yilmaz, 2021)

The suggested method's primary contribution is the demonstration of a hybrid model developed by effectively combining two well-known, previously trained network models. Two different types of datasets were used to assess the suggested deep learning technique. Here, a comparison between the proposed hybrid model and each individual model was initially made. The test results demonstrated that the suggested approach can successfully classify malware with respect to f-score, recall, accuracy and precision. Furthermore, it is noted that the suggested approach is effective in minimizing feature space across a broad domain (Aslan & Yilmaz, 2021).

In this paper (Aslan, Ozkan-Okay & Gupta, 2021), authors present an intelligent system for detecting malware on cloud computing based on the behavior of the user. In the proposed system, a suspicious file is uploaded by the user via a computer network to the cloud environment. The given file is then run via several virtual machines (VMs), and the necessary dynamic tools are used to collect the execution traces. On a behavior-based detection agent, generated behaviors are gathered from generated execution traces. To construct features, similar actions are grouped based on predetermined rules. A suggested cloud-based behavior centric model (CBCM) is utilized for developing features. Following that, the majority of discriminative characteristics are chosen using recommended techniques, and the features that make the cut are transmitted to the detection agents, which include rule- and learning-based detection. Selective characteristics are taught using machine learning algorithms, such as logistic model trees (LMT), C4.5 (J48), random forest (RF), sequential minimum optimization (SMO), k-nearest neighbor (KNN), and simple logistic regression (SLR), in learning-based detection agents. In contrast, features in rule-based detection agents are assessed using predetermined feature sets. Every sample is classified as either benign or malicious and saved in the database using detection agents that are based on rules and learning. The user receives the analysis result back, which indicates whether or not the suspected file contains malware. The authors tested the methods on custom datasets, including normal and malicious files.

Javaheri, Hosseinzadeh & Rahmani, (2018) present a revolutionary approach to identify, track, and combat ransomware and other stealthy malware, such as keyloggers, screen recorders, and blockers. This paper's suggested approach is based on an exhaustive and transparent hooking of kernel-level procedures that allows for a dynamic behavioral analysis. Three malware classifications were identified through the application of the JRIP, J48 decision tree, and linear regression techniques as classifiers. The primary architectural design of an anti-spyware program is presented in this paper. It tracks spyware footprints to identify and stop processes that are in use, remove executable files, and limit network traffic. The effectiveness of the suggested approach was assessed from the perspectives of accuracy in ROC curve analysis for real-world spyware samples and success rate in effectively confronting active spyware. The suggested approach had an error rate of about 7% and an accuracy of almost 93% when it came to identifying malware. Furthermore, the proposed method has an approximately 82% success rate in removing spyware from an operating system. The method has used a dataset containing 4951 real-world samples of spyware and 3025 benign executable files.

While the security challenge under malevolent opponents has received little attention, the topic of learning-based command for robots has been thoroughly studied. Robots' intelligent gadgets and communication networks are vulnerable to invasions by malicious opponents, which can result in mishaps, accomplish unlawful goals, and even cause personal injury. In the context of deep reinforcement learning, the article (Wu et al., 2023) first looks into the issues of designing countermeasures and scheduling the best false data injection attacks for robots that resemble cars. An optimal fake data injection attack technique is suggested to degrade a robot's tracking performance by assuring the tradeoff between attack efficiency and restricted attack energy, using an innovative deep reinforcement learning approach. After that, the best tracking control method is discovered to lessen attacks and restore tracking performance. More significantly, a robot employing the learning-based secure command approach is guaranteed to be stable theoretically. The usefulness of the proposed systems is demonstrated through a combination of simulated and real-world trials.

Because of insecure wireless communication, resource-constrained architecture, a variety of IoT device types, and a large volume of sensor data being transferred over the network, networks enabled by the Internet of Things (IoT) are extremely susceptible to cyber threats. IoT-compatible security solutions are therefore necessary. One of the most popular methods for identifying cyberthreats in IoT-enabled networks is the use of intrusion detection systems. Unfortunately, several problems plague the majority of the current cyber threat detection technologies, including inadequate scalability, high false positive rate (FPR), poor accuracy, and excessive learning complexity. To address these problems, the authors of the article (Dey, Gupta & Sahu, 2023) suggest a unique and intelligent metaheuristic-based framework for cyber threat detection that makes use of ensemble selection and categorization of features techniques. In order to obtain an optimum collection of features and prevent the curse of dimensionality for effective learning, a metaheuristic-based ensemble feature selection framework is first created utilizing the Binary Gravitational Search Algorithm (BGSA) and Binary Grey Wolf Optimization (BGWO). Next, to identify and categorize cyber threats, Decision Tree and ensemble learning-based classification approaches like AdaBoost and Random Forest (RF) are used independently. The RF beats current modern cyber threat detection methods, according to the result analysis, because of its improved feature subset (4 features out of 42), maximum accuracy (99.41%), maximum detection rate (99.09%), maximum F1-score (99.33%), and lowest FPR (0.03%). The proposed work used the UNSW-NB15 dataset for assesing the efficiency of the proposed experiments.

The cornerstone of advanced metering infrastructure (AMI) is bi-directional communication networks, yet these networks also put smart grids at significant danger of intrusion. Although a number of intrusion detection systems (IDS) for AMI have been developed in earlier research, most of them have not fully taken into account the influence of various parameters on incursions. This research (Yao et al., 2022) suggests an intrusion detection system (IDS) based on deep learning theory to protect the bi-directional communication network of the AMI. The adaptive synthetic (ADASYN) sampling method balances the data distribution after the invalid features are removed first using a characteristic screening approach based on eXtreme Gradient Boosting (XGBoost). To improve the spatial distribution of the data, multi-space feature subsets built using a convolutional neural network (CNN) are then created. The proposed system was tested using the KDDCup99, NSL-KDD, and CICIDS-2017 datasets.

Yamin et al., (2024) present a novel method for creating dynamic, intricate, and flexible cybersecurity exercise situations using Large Language Models (LLMs). Inspired by Turing's seminal investigation into machine cognition, that raises doubts about computers' capacity to replicate intellect and thought in humans. The approach enhances cybersecurity awareness and training by simulating a variety of known and unknown cyber dangers by taking advantage of the generative power of LLMs. By using this technique, the "hallucination" potential that exists in LLMs can be turned into a potential benefit, allowing for the development of challenging exercise situations that go beyond the parameters of conventional cybersecurity training. The creative use of AI to improve security personnel’s readiness for various cyberthreats is where the innovation resides. To guarantee their realism and applicability, the scenarios created using this technology underwent extensive validation and a rigorous evaluation procedure that included expert review, GPT models, and expert assessment. In addition, the prompts given to the LLMs were carefully crafted to utilize a Retrieval-Augmented Generation (RAG) methodology, enhancing the intricacy and significance of the situations. This integration of RAG demonstrates a sophisticated use of AI in cybersecurity instruction, revealing a profound knowledge of how machines can enhance capacities to foresee and mitigate cyber risks. It also draws inspiration from Turing's investigation of machine intelligence.

Quinn & Thompson, (2024) present a novel method for creating and putting into practice cybersecurity policies intended to mitigate spear phishing attempts against senior corporate managers by utilizing Google Gemini's generative AI capabilities. The study showed that by fusing cutting-edge artificial intelligence with established security standards, corporate cybersecurity frameworks might significantly improve their identification, prevention, and response tactics. In addition to enabling dynamic policy modifications based on real-time data analysis, the deployment of machine learning algorithms enhanced threat detection speed and accuracy. This has proven vital in the ever-evolving world of digital threats. The results highlight how AI has the ability to revolutionize cybersecurity procedures by providing more flexible, proactive, and strong protections against ever-more-advanced spear phishing tactics. The research delves deeper into the consequences of AI-driven regulations on business governance and compliance, proposing a novel framework wherein AI not only facilitates but also actively shapes strategic security determinations. The encouraging outcomes suggest that further research should be done on the wider uses of artificial intelligence in cybersecurity and hint to a time when integrating AI into protection strategies against sophisticated cyberattacks would be commonplace. The datasets sources were NIST 800 - 61 for incident response, COBIT 2019 guidelines on phishing management, ISO standards for managing IT risks, and various government policies that focus on identity management and cyber threat mitigation.

In this paper (Long et al., 2024), the authors evaluated the feasibility and predicted accuracy of a Transformer-based network intrusion detection method designed for cloud environments. In Figure 2 the architecture of the model can be seen, which is divided into three stages: data processing, model training and prediction and model evaluation.

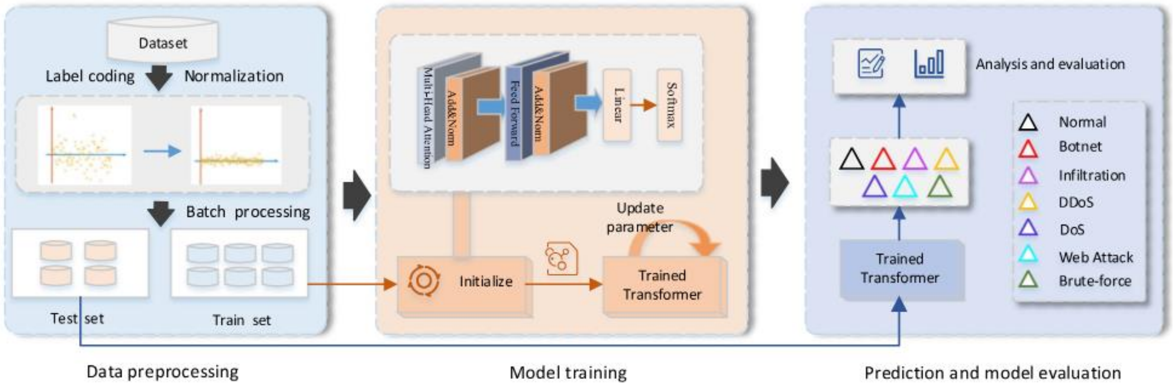

Figure 2. The architecture of Transformer based network intrusion detection model (Long et al., 2024)

They created and put into practice the proposed algorithm by utilizing the Transformer model's attention mechanism in conjunction with network intrusion detection concepts. Preprocessing the dataset, training the model with three encoder layers at first, and assessing the model's predictive ability were the steps in the process. Then, in order to train the model further, gradually added more encoder layers, and compared the model to the well-known CNNLSTM model. The final findings demonstrate the effectiveness of the Transformer-based network intrusion detection model in predicting network intrusions within cloud environments. The method achieved a prediction accuracy of over 93% under the specified experimental conditions, which is comparable to the most recent approach on the CNN-LSTM model (Long et al., 2024).

Malware has evolved in recent years by employing various obfuscation tactics; as a result, malware detection has grown more difficult. Malware detectors that rely on signatures or conventional behavior are unable to identify this new breed of malware. In order to generate a malware dataset that captures semantically linked behaviors from sample programs, the authors of the paper (Aslan, Samet & Tanrıöver, 2020) depicted a subtractive center behavior model (SCBM). The suggested model takes into account system routes, the locations of virus behaviors, and the behaviors themselves. In this manner, patterns of harmful conduct are distinguished from patterns of benign behavior. Features that did not meet the predetermined threshold for scoring are eliminated from the dataset. Compared to datasets generated using n-gram and other models that have been utilized in earlier studies, the datasets constructed using the suggested model have far fewer features. Both known and new malware can be handled by the proposed model, which also outperforms established models in terms of accuracy and detection rate. Two datasets – one with a score and the other without – are generated using SCBM in order to demonstrate the efficacy of the model. 3000 benign samples and 6700 malware samples were examined in total. The outcomes were contrasted with models from other studies in literature and those obtained using n-gram. The test results demonstrated that the accuracy, false positive rate, and detection rate were assessed at

99.8%, 0.2%, and 99.9%, respectively, when the suggested model was combined with a suitable machine learning method.

Jayaraman et al., (2023), designed a system to identify IoT device communications as either benign or malicious by utilizing traffic packets containing essential information as features for pattern identification and classification. Both machine-learning and deep-learning classifiers are employed for IoT traffic classification, with Decision Tree, Random Forest, SVM, and Convolutional Neural Network (CNN) serving as binary classifiers to predict communication nature. The IoT communication data undergoes several pre-processing stages to generate and prepare the salient feature vectors, which are then processed by the four classifiers. The most effective classifier for identifying secure IoT communications is determined by evaluating the models' performance using various evaluation measures. Using the IoT-23 dataset, the proposed machine learning approach achieves an accuracy of 99.25%.

A detailed comparison of AI models is presented in the paper (Floroiu et al., 2024) to explore methods for detecting malware, specifically Remote Access Trojans (RATs). To enhance detection capabilities, the research utilizes a dataset generated by computing texture images from a substantial number of executable files infected with RATs. The efficacy of various machine learning models – namely VGG-16, VGG-19, and ResNet50 – is thoroughly investigated in identifying these malicious programs. Additionally, the study explores transformer architectures, such as the Vision Transformer (ViT), which are particularly adept at image classification tasks. The accuracy of the models is reported in the following order: VGG-16 (69.81%), VGG-19 (73.47%), ResNet50 (76.83%), and Vision Transformer (83.26%).

The integration of AI in cybersecurity represents a significant evolution in the field, providing advanced tools and techniques to combat modern threats. The continuous development of AI applications not only improves the efficacy of cybersecurity measures but also ensures that organizations can proactively defend against an ever-changing threat landscape. Through enhanced threat detection, automated responses, robust data protection, and comprehensive network security, AI stands as a cornerstone for the future of cybersecurity.

3. Comparative analysis

Building upon the AI applications and techniques discussed in the preceding chapter, Table 1 presents a comprehensive comparative analysis that illustrates the accuracy of each method. This analysis reveals that models based on behavior analysis exhibit the highest accuracy.

Table 1. Comparative analysis of the accuracy of the method presented

Article	Accuracy
Aslan, O. & Yilmaz, A.A. (2021) A New Malware Classification Framework Based on Deep Learning Algorithms.	97.78%
Aslan, O., Ozkan-Okay, M. & Gupta, D. (2021) Intelligent Behavior-Based Malware Detection System on Cloud Computing Environment.	99.7%
Javaheri, D., Hosseinzadeh, M. & Rahmani, A.M. (2018) Detection and Elimination of Spyware and Ransomware by Intercepting Kernel-Level System Routines.	93%
Dey, A.K., Gupta, G.P. & Sahu, S.P. (2023) A metaheuristic-based ensemble feature selection framework for cyber threat detection in IoT-enabled networks	99.41%
Yao, R., Wang, N., Chen, P., Ma, D. & Sheng, X. (2022) A CNN-transformer hybrid approach for an intrusion detection system in advanced metering infrastructure.	97.85%
Yamin M., Hashmi E., Ullah M., Katt B. (2024) Applications of LLMs for Generating Cyber Security Exercise Scenarios.	Not available
Quinn T., Thompson O. (2024) Applying Large Language Model (LLM) for Developing Cybersecurity Policies to Counteract Spear Phishing Attacks on Senior Corporate Managers.	25%

Long, Z., Yan, H., Shen, G., Zhang, X., He, H. & Cheng, L. (2024) A Transformer-based network intrusion detection approach for cloud security.	93%
Aslan, Ö., Samet, R. & Tanrıöver, Ö.Ö. (2020) Using a Subtractive Center Behavioral Model to Detect Malware.	99.8%
Bhuvana Jayaraman, Mirnalinee Thanga Nadar Thanga Thai, Anirudh Anand, Karthik Raja Anandan, „Detecting malicious IoT traffic using Machine Learning techniques.	99.25%
Iustin Floroiu, Miruna Floroiu, Alexandru-Constantin Niga, Daniela Timisica, "Remote Access Trojans Detection Using Convolutional and Transformer-based Deep Learning Techniques”.	83.26%

4. Integrating artificial intelligence in the cyber security field

In the paper (Kim & Park, 2024), several key domains are identified where AI applications can significantly enhance cybersecurity efforts. These domains include:

- Intrusion and threat detection;
- Security monitoring;
- Vulnerability scanning and remediation;
- Data classification;
- Spam filtering and social engineering detection;
- Security automation;
- User behavior analytics;
- Network traffic profiling and anomaly detection.

These areas demonstrate the diverse and impactful ways AI can be leveraged to strengthen cybersecurity measures across various aspects of digital security.

In the article (Kaur et al., 2023), the authors provide a comprehensive categorization of the most critical pillars of cybersecurity and identify a wide range of sectors where AI applications can significantly enhance essential processes. The selection includes the general cyber security areas mentioned earlier. Table 2 outlines these pillars of cybersecurity, the key sectors, and the specific AI applications that can be integrated:

Table 2. Visual representation of the cybersecurity pillars, sectors of activity and AI applications

Pillar of Cybersecurity	**Sector**	**AI Application**
Identify	Asset Management	• asset inventory management • automated configuration management • automated security control validation
	Business Environment	• automated business impact analysis
	Governance	• automated policy enforcement
	Risk Assessment	• automated vulnerability identification and assessment • predictive intelligence • attack path modelling • automated threat hunting • automated risk analysis and impact assessment
	Risk Management Strategy	• decision support for risk planning
	Supply Chain Risk Management	• cyber supply chain security

Protect	Identify Management, Authentification and Acess Control	• AI-supported user authentification • AI- supported device authentification • automated access control
	Awareness and Training	• adaptive security awareness and training
	Data Security	• data leakage prevention • intelligent e-mail protection • malicious domain blocking and reporting
	Information Protection, Process and Technology	• AI-powered backup • AI-enhanced vulnerability management plan
	Protective Technology	• log analysis • protection by deception • anti-virus/anti-malware • intrusion prevention system
Detect	Anomalies and Events	• intrusion detection system (IDS)
	Security Continuous Monitoring	• security monitoring
	Detection Processes	• automated assessment of threat intelligence sources • multi-lingual threat intelligence • dark web investigation • AI-powered honeypots
Respond	Communications	• automated responsibility allocation • collaboration support system
	Analysis and Respond Planning	• alert triage • forensic analysis • automated incident characterization
	Mitigation	• automated isolation • automated remediation
	Improvements	• long-term improvements
Recover	Recovery Planning	• automated recommendation platform
	Improvements	• analysis and aggregation of incidents reports
	Communication	• information sharing propriety platform

4.1. Advantages of using AI in cybersecurity activitites

In recent years, conflicts have increasingly been waged in cyberspace, and with the continuous evolution of cyber attacks and threats, traditional cybersecurity techniques and methods are no longer sufficient to address these challenges. Artificial Intelligence (AI) offers significant advantages over traditional cybersecurity approaches, providing innovative ways to combat modern threats. The main advantages of integrating AI into the field of cybersecurity include (Jawaid, 2023):

- **Efficiency and scalability**: it processes and analyzes vast volumes of data far more quickly and accurately than human analysts. As a result, security specialists have less work to do and may focus on harder and more complex projects. Furthermore, because AI-powered security solutions can analyze data from multiple sources and handle massive traffic volumes without appreciably reducing performance, they are very scalable.Thus, there are several advantages of using AI in cybersecurity. For the digital age, its effectiveness, scalability, perpetually learning, adaptive capabilities, speed and accuracy in identifying and responding to threats make it an indispensable instrument for cyberattack defense.
- **Speed and Accuracy**: the potential of artificial intelligence to provide real-time threat recognition and response is one of its key advantages for cybersecurity. AI algorithms can swiftly detect any unexpected behavior that can point to a security risk by looking at network traffic. As a result, security officers can stop the attack before it does too much harm. AI-powered security systems are also more precise and reliable than conventional ones since they

can evaluate enormous volumes of data and spot possible threats that human analysts could have overlooked.

- **Continuous Learning and Adaptation**: the capacity of AI to develop and learn over time is a key benefit for cyber security. Artificial Intelligence (AI) algorithms may detect patterns and generate insights that increase their efficacy and accuracy by assessing security occurrences and continuously monitoring security systems. AI also makes it easier to keep one step ahead of hackers by enabling security systems to respond promptly to emerging risks and attack methods.

By leveraging these AI capabilities, organizations can significantly bolster their cybersecurity defenses, making them more resilient against the sophisticated threats of today’s digital landscape.

4.2. Challenges in adopting AI in Cybersecurity

The implementation of AI methods and algorithms in cybersecurity might provide a number of challenges as with any novel approach or integration (Sontan & Samuel, 2024):

- **Quantity and Quality of Data**: For artificial intelligence (AI) systems to work well, a lot of high-quality data must be collected. Inadequate or poor quality data can result in imprecise forecasts and inefficient security protocols.
- **Algorithm bias**: Algorithms using artificial intelligence may carry over biases from the training set, producing biased results. This can be especially troublesome in applications related to security where accuracy and fairness are essential.
- **Adversarial Attacks**: Malevolent actors may alter inputs to trick artificial intelligence (AI) systems, causing the AI to make false assessments of risk or fail to detect threats altogether.
- **Complexity and Cost**: Putting AI systems into place and keeping them maintained can be costly and complex. To properly implement AI, organizations need certain knowledge and resources.
- **Integration with Current Systems**: It can be difficult to integrate AI technologies with the current cybersecurity framework. It takes careful design and implementation to ensure compatibility and seamless functioning alongside traditional systems.
- **AI systems must always learn new things and adjust to new threats**. This requires continuous learning and adaptation. In order to guarantee that the AI models continue to be successful against changing cyberthreats, continuous monitoring and updating is necessary.

In order to create solid, dependable and equitable security solutions, as well as to fully reap the positive effects of AI in cybersecurity, these issues must be resolved.

5. Ethical concerns

The adoption of artificial intelligence (AI) raises multiple challenges, particularly on the ethical front. While it is true that AI can enhance cybersecurity methods, it also has the potential to introduce new vulnerabilities and risks (Adeniyi & Ness, 2024):

- **Bias and Fairness:** AI algorithms can inadvertently reflect biases present in their training data, leading to discriminatory outcomes. This unintentional bias may result in unfair treatment, especially in critical areas such as threat detection, access control, and incident response. Such biases can cause disproportionate suspicion towards certain groups, unjustly restrict access for specific individuals, and skew incident response priorities, undermining the fairness and effectiveness of AI-driven cybersecurity measures.
- **Transparency and Explainability:** AI algorithms, particularly deep learning models, are often seen as "black boxes" due to their complexity and difficulty in interpretation. This lack of

transparency and explainability raises significant concerns about accountability. It also challenges the ability to understand the reasoning behind AI-driven cybersecurity decisions, making it difficult to verify and trust these systems.

- **Accountability and Responsibility:** Determining who is responsible for mistakes made by AI systems or for failing to stop security incidents is a big problem. The inability to assign precise blame might make it more difficult to address and correct mistakes. Due to the difficulty in determining the individual accountable for the system's shortcomings, this ambiguity might give rise to ethical and legal difficulties.
- **Privacy Concerns:** Large volumes of personal data may be processed by AI systems for threat analysis and detection. This prompts worries about possible privacy violations, especially if artificial intelligence algorithms don't have strong privacy safeguards. When personal data is used without sufficient protections, ethical concerns about use of information and consent may arise.
- **Overreliance on AI:** One risk of overestimating AI's potential is creating an illusion of security. Over-reliance on AI in the absence of sufficient human oversight could lead to vulnerabilities, false positives, and threats that AI is unable to identify or mitigate. To ensure complete cybersecurity, it is imperative to strike a balance between AI integration and ongoing human monitoring.
- **Adversarial Attacks:** Adversaries may try to introduce false information into AI systems in an effort to manipulate them. This makes one wonder how resilient AI models are to these kinds of hostile attacks. Attacks against AI systems have the potential to weaken cybersecurity defenses and jeopardize the system's integrity if they are not sufficiently safeguarded.
- **Data Security and Misuse:** Data security is essential for both training and using AI models. This data is susceptible to breaches that could allow for unauthorized access and possible misuse if it is not adequately secured. For AI systems to remain reliable and honest, strong data protection protocols must be in place.
- **Lack of Standards and Regulations:** One major problem is the lack of uniform ethical standards and laws pertaining to AI in cybersecurity. Inconsistent ethical behavior and the abuse of AI technologies are risks associated with the absence of established rules. To guarantee the ethical and responsible application of AI in the cybersecurity space, consistent criteria must be established.
- **Long-Term Implications:** There are important concerns regarding the long-term social effects of extensive AI use in cybersecurity. Concerns include how people will play a changing role in security, how much society trusts AI systems, and how this may affect democratic values more broadly. A multidisciplinary strategy combining engineers, ethicists, legislators and other stakeholders is necessary to address these ethical challenges. This entails creating unambiguous ethical standards, encouraging accountability, openness, and making sure AI technologies are created and used with meticulous consideration for their wider society impact.

To tackle these concerns, a multidisciplinary strategy incorporating transparent procedures, strong privacy safeguards, ethical standards and accountability mechanisms is required. Ensuring that AI technologies enhance cybersecurity while adhering to moral principles and societal norms is achievable by considering these factors.

6. Conclusions

In conclusion, there are both opportunities and difficulties in the rapidly changing fields of cybersecurity and artificial intelligence. Artificial intelligence (AI) is becoming an indispensable weapon in the fight against more complex cyberattacks due to its capacity to change, grow and analyze large amounts of information. But while using it, it's critical to understand its limitations and take ethical considerations into account. Potential biases, a lack of transparency, problems with accountability, and intrusions of privacy are some of these worries. It will take a multidisciplinary approach incorporating engineers, ethicists, policymakers, and various other actors to ensure that AI systems are designed and used responsibly. The combination of artificial intelligence and cybersecurity will become more and more important in protecting people, businesses, and society at large as the digital world develops. The defenses against cyber enemies can be strengthened by the incorporation of AI, which can improve threat detection, automate reaction mechanisms, and provide better data protection measures. In addition, the cybersecurity landscape is expected to undergo a radical change with the introduction of quantum computing and other cutting-edge technology. Due to the unmatched processing power of quantum computing, conventional encryption approaches may be compromised, hence requiring the creation of quantum-resistant cryptography algorithms. This transformation will bring new issues that need to be overcome as well as new opportunities to improve cybersecurity. All things considered, the future of cybersecurity rests on the ethical and cautious integration of AI, combined with ongoing innovation to combat the new risks posed by cutting edge technologies such as quantum computing. It can be guaranteed that the advantages of AI along with other cutting-edge technologies will be achieved while protecting against their potential threats and upholding public trust by promoting openness, accountability, and strong ethical standards.

REFERENCES

Adeniyi, S. & Ness, S. (2024) The Role of Artificial Intelligence in Cybersecurity, Researchgate [online], Available at: https://www.researchgate.net/publication/377223280_The_Role_of_Artificial_Intelligence_in_Cybersecurity_Authors

Aslam, M. (2024) AI and Cybersecurity: An Ever-Evolving Landscape. *International Journal of Advanced Engineering Technologies and Innovations*. 1(1), 52–71. doi.org/10.765656/9jcnq589.

Aslan, O., Ozkan-Okay, M. & Gupta, D. (2021) Intelligent Behavior-Based Malware Detection System on Cloud Computing Environment. In *IEEE Access*. 9, 83252–83271. doi:10.1109/access.2021.3087316.

Aslan, O. & Yilmaz, A.A. (2021) A New Malware Classification Framework Based on Deep Learning Algorithms. In *IEEE Access*. 9, 87936–87951. doi:10.1109/access.2021.3089586.

Aslan, Ö., Samet, R. & Tanrıöver, Ö.Ö. (2020) Using a Subtractive Center Behavioral Model to Detect Malware. *Security and Communication Networks*. 2020(1), 1–17. doi: 10.1155/2020/7501894.

Jayaraman, B., Thanga Nadar Thanga Thai, M., Anand, A. & Anandan, K.R. (2023) Detecting malicious IoT traffic using Machine Learning techniques. *Romanian Journal of Information Technology and Automatic Control.* 33(4), 47-58. doi:10.33436/v33i4y202304.

Dey, A.K., Gupta, G.P. & Sahu, S.P. (2023) A metaheuristic-based ensemble feature selection framework for cyber threat detection in IoT-enabled networks. *Decision Analytics Journal*. 7(100206), 1-16. doi:10.1016/j.dajour.2023.100206.

Floroiu, I., Floroiu, M., Niga, A.-C. & Timisica, D. (2024) Remote Access Trojans Detection Using Convolutional and Transformer-based Deep Learning Techniques. *Romanian Cyber Security Journal*. 6(1), 47-58. doi:10.54851/v6i1y202405.

Javaheri, D., Hosseinzadeh, M. & Rahmani, A.M. (2018) Detection and Elimination of Spyware and Ransomware by Intercepting Kernel-Level System Routines. In *IEEE Access*. 6, 78321–78332. doi: 10.1109/access.2018.2884964.

Jawaid, S.A. (2023) Artificial Intelligence with Respect to Cyber Security. *Journal of Advances in Artificial Intelligence.* 1(2), 96-102. doi: 10.20944/preprints202304.0923.

Kaur, R., Gabrijelčič, D. & Klobučar, T. (2023) Artificial Intelligence for Cybersecurity: Literature Review and Future Research Directions. *Information Fusion*. 97(101804), 1-29. doi: 10.1016/j.inffus.2023.101804.

Kim, G. & Park, K. (2024) Effect of AI: The Future Landscape of National Cybersecurity Strategies.*Technical Journal (Tehnički Glasnik)*. 18(1), 29–36. doi: 10.31803/tg-20230218142012.

Kumar, S., Gupta, U., Singh, A. & Singh, A.K. (2023) Artificial Intelligence. *Journal of Computers Mechanical and Management*. 2(3), 31–42. doi: 10.57159/gadl.jcmm.2.3.23064.

Long, Z., Yan, H., Shen, G., Zhang, X., He, H. & Cheng, L. (2024) A Transformer-based network intrusion detection approach for cloud security. *Journal of Cloud Computing*. 13(5). doi: 10.1186/s13677-023-00574-9.

Quinn T. & Thompson O. (2024) Applying Large Language Model (LLM) for Developing Cybersecurity Policies to Counteract Spear Phishing Attacks on Senior Corporate Managers. To be published in *Research Square*. [Preprint (Version 1)] https://doi.org/10.21203/rs.3.rs-4405206/v1 [Accessed May 2024].

Sontan, A.D. & Samuel, S.V. (2024) The intersection of Artificial Intelligence and cybersecurity: Challenges and opportunities. *World Journal Of Advanced Research and Reviews*. 21(2), 1720–1736. doi: 10.30574/wjarr.2024.21.2.0607.

Yamin, M., Hashmi, E., Ullah, M., & Katt, B. (2024) Applications of LLMs for Generating Cyber Security Exercise Scenarios. To be published in *Research Square*. [Preprint (Version 1)] https://doi.org/10.21203/rs.3.rs-3970015/v1 [Accessed May 2024].

Yao, R., Wang, N., Chen, P., Ma, D. & Sheng, X. (2022) A CNN-transformer hybrid approach for an intrusion detection system in advanced metering infrastructure. *Multimedia Tools and Applications*. 82(13), 19463-19486. doi:/10.1007/s11042-022-14121-2.

Wu, C., Yao, W., Luo, W., Pan, W., Sun, G., Xie, H. & Wu, L. (2023). A Secure Robot Learning Framework for Cyber Attack Scheduling and Countermeasure. In *IEEE Transactions on Robotics*. 39(5), 3722–3738. doi: 10.1109/tro.2023.3275875.

Andreea DINU is a Scientific Researcher within the Cybersecurity Department of the National Institute for Research and Development in Informatics – ICI Bucharest. With a strong technical background, Andreea has actively participated in national and international research projects across various domains, including cybersecurity, artificial intelligence, high-performance computing and quantum computing. Andreea is a dedicated researcher with numerous publications that highlight her significant contributions to these fields. Additionally, she is a member of the technical group of the European Blockchain Partnership, where she contributes to the development and implementation of distributed ledger technologies within the European Union. Her involvement in such projects demonstrates her commitment to technological advancement and digital security, ensuring that ICI Bucharest remains at the forefront of research and innovation in informatics.

Andreea DINU este Cercetător Științific în cadrul Departamentului de Securitate Cibernetică al Institutului Național de Cercetare-Dezvoltare în Informatică – ICI București. Cu un background tehnic solid, participă activ la proiecte de cercetare naționale și internaționale în diverse domenii, inclusiv securitate cibernetică, inteligență artificială, calcul de înaltă performanță și calcul cuantic. Andreea Dinu este o cercetătoare dedicată, cu numeroase publicații care

evidențiază contribuțiile sale semnificative în aceste domenii. De asemenea, este membră a grupului de lucru tehnic al Parteneriatului European pentru Blockchain (EBP), unde contribuie la dezvoltarea și implementarea tehnologiilor de registru distribuit în cadrul Uniunii Europene. Implicarea ei în astfel de proiecte demonstrează angajamentul față de avansarea tehnologică și securitatea digitală, asigurând că ICI București rămâne în avangarda cercetării și inovației în domeniul informaticii.

Paul Cristian VASILE is a Scientific Researcher and Chief of Cyber Security Department at the National Institute for Research and Development in Informatics – ICI Bucharest. With a robust background in computer science and cybersecurity, he has significantly contributed to the institute's mission of advancing technological innovation and safeguarding critical infrastructures. Paul is actively involved in numerous collaborative projects at both the national and international levels and his dedication to the cybersecurity field is also reflected in his numerous publications and presentations at prestigious conferences. His technical background allowed him to contribute to cybersecurity strategy development, threat analysis, and the implementation of advanced security protocols.

Paul Cristian VASILE este cercetător științific și șef al Departamentului de Securitate Cibernetică la Institutul Național de Cercetare-Dezvoltare în Informatică – ICI București. Cu un background solid în informatică și securitate cibernetică, el a contribuit semnificativ la misiunea Institutului de a avansa inovația tehnologică și de a proteja infrastructurile critice. Paul Crisitan Vasile este activ implicat în numeroase proiecte de colaborare atât la nivel național, cât și internațional, iar dedicarea sa în domeniul securității cibernetice se reflectă și în numeroasele sale publicații științifice și prezentări susținute la conferințe prestigioase. Expertiza sa tehnică i-a permis să contribuie la dezvoltarea strategiilor de securitate cibernetică, analiza amenințărilor și implementarea protocoalelor de securitate avansate.

Alexandru GEORGESCU is a Scientific Researcher II within the Department for Cybersecurity and Critical Infrastructure Protection of the National Institute for Research and Development in Informatics - ICI Bucharest. He has an eclectic background, having studied Economics, then Geopolitics, and has obtained a Ph.D. in Risk Engineering for Critical Infrastructure Systems with a thesis on Critical Space Infrastructures, which was expanded into a book and published by Springer in 2019. He is actively involved in advancing Critical Infrastructure Protection and Resilience issues through cooperation at international level and has worked on international projects for the European Space Agency and others. Since 2019, he is a co-moderator of the Working Group on the Protection of Defence-related Critical Energy

Infrastructures within the Consultation Forum on Sustainable Energy in Defence and Security Sectors organized by the European Defence Agency. In this position, he has contributed to policy documents, project proposals and studies on Critical Infrastructure Protection, including from the perspective of cyber resilience and hybrid threats. He is affiliated with the Romanian Association for the Promotion of Critical Infrastructure Protection, with the Romanian Association for Space Technology and Industry and Eurodefense Romania.

Alexandru GEORGESCU este cercetător științific gradul II în cadrul Departamentului de Securitate Cibernetică și Protecția Infrastructurii Critice al Institutului Național de Cercetare-Dezvoltare în Informatică – ICI București. Are un background eclectic, studiind inițial Economie, apoi Geopolitică, și obținând un doctorat în Ingineria Riscului pentru Sisteme de Infrastructură Critică, pe tema despre Infrastructurile Spațiale Critice. Teza sa a fost extinsă într-o carte publicată de Springer în 2019. De asemenea, contribuie activ la soluționarea problemelor de Protecție și Reziliență a Infrastructurii Critice prin colaborare internațională și a fost implicat în proiecte internaționale pentru Agenția Spațială Europeană și alte organizații. Din anul 2019, este co-moderator al Grupului de Lucru 3 pe Protecția Infrastructurilor Critice de Energie, în cadrul Forumului Consultativ pentru Energie Sustenabilă în Sectoarele de Apărare și Securitate (CF SEDSS), organizat de Agenția Europeană de Apărare (EDA). În această poziție, a contribuit la elaborarea documentelor de politici, propunerilor de proiecte și studiilor privind Protecția Infrastructurii Critice, cu un accent deosebit pe reziliența cibernetică și amenințările hibride. Este afiliat la Asociația Română pentru Promovarea Protecției Infrastructurii Critice (APCS), cu Asociația Română pentru Tehnologie și Industrie Spațială (ROMSPACE) și la Eurodefense România.