International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

Int. J. Sci. Res. Arch.
IJSRA
International Journal of Science and Research Archive
Research Journal Archive, INDIA

(Research Article)

Challenges and ethical implications of using AI in cybersecurity

Oluwafemi S. Ajibola [1, *], Oladipupo Dopamu [1] and Olawunmi Olurin [2]

[1] Department of Computer Science, Western Illinois University, United States of America.
[2] Department of Information Technology, American National University, Salem, Virginia, USA.

International Journal of Science and Research Archive, 2025, 14(02), 294-304

Publication history: Received on 18 December 2024; revised on 27 January 2025; accepted on 30 January 2025

Article DOI: https://doi.org/10.30574/ijsra.2025.14.2.0276

Abstract

Integrating Artificial Intelligence (AI) into cybersecurity has transformed the field by enhancing threat detection, automating responses, and predicting vulnerabilities. AI-driven tools such as intrusion detection systems, malware analysis, and user authentication mechanisms have significantly improved efficiency and accuracy. However, adopting AI also presents challenges, including data biases, adversarial attacks, high computational demands, and ethical concerns such as privacy violations and dual-use problems. This paper examines the applications, challenges, and ethical implications of using AI in cybersecurity, providing insights into its limitations and strategies for responsible deployment. It emphasizes the importance of balancing innovation with ethical considerations to ensure the effectiveness and fairness of AI in mitigating cyber threats.

Keywords: Artificial Intelligence; Cybersecurity; Ethical Implication; Privacy; Cyber-Attacks; Cyber Threats

1. Introduction

Artificial Intelligence (AI) enables machines to mimic human behavior and intelligence efficiently. It encompasses subfields such as machine learning (ML), which involves training algorithms to learn from data; deep learning (DL), a subset of ML focused on neural networks with multiple layers; federated learning (FL), a decentralized approach to training models across distributed devices while preserving data privacy; and explainable AI (XAI), which aims to make AI systems more transparent and interpretable. AI applications span various sectors, including healthcare [1,2], manufacturing [3], education [4], communication networks [5,6], agriculture [7], power grids [8], finance [9], and government [10]. This rapid adoption is changing the way these industries operate, with AI projected to contribute up to $13 trillion to the global economy by 2030, increasing the global gross domestic product (GDP) by 26% [11]. Among its transformative applications, AI has achieved significant successes in cybersecurity, from enhancing threat detection to automating responses and predicting vulnerabilities.

Cybersecurity is the practice of protecting systems, networks, and data from cyber threats, ensuring confidentiality, integrity, and availability in the digital domain. It encompasses a range of technologies, strategies, and practices to mitigate risks and safeguard critical assets [12]. The application of AI in cybersecurity has evolved from simple rule-based systems to advanced ML models capable of detecting complex cyber threats. Early implementations focused on automating routine tasks, such as filtering spam emails and detecting known malware signatures. With the advent of ML and DL, AI systems have gained the ability to detect previously unknown threats by analyzing patterns and anomalies in large datasets. For instance, ML algorithms can classify new malware based on behavioral analysis, while natural language processing techniques extract actionable insights from threat intelligence reports. Recent advancements have further enabled real-time threat response and the prediction of potential vulnerabilities, significantly improving the security posture of cybersecurity systems.

[*] Corresponding author: Oluwafemi S. Ajibola.

Despite these advancements, the integration of AI into cybersecurity is not without challenges. AI systems are susceptible to adversarial attacks, where malicious actors manipulate inputs to deceive the model, potentially leading to security breaches. Moreover, the reliance on large datasets for training raises concerns about the ethical implications of using AI in cybersecurity. This includes data privacy, when AI systems process vast amounts of personal and sensitive data, raising questions about data protection and user consent, as well as bias, when AI systems are trained using a particular set of data, potentially leading to unfair or discriminatory practices. The lack of transparency in AI decision-making processes, often referred to as the "black box" problem, further complicates its adoption in critical security systems. Based on these, this paper explores the challenges and ethical implications of using AI in cybersecurity. By examining current applications, identifying key challenges, and discussing ethical considerations, this study aims to provide insights that can guide the responsible development and deployment of AI technologies in the field of cybersecurity. Balancing innovation with ethical responsibility is crucial to ensuring that AI serves as a force for good in the fight against cyber threats.

The rest of the paper is organized as follows: Section 2 presents and reviews the specific applications of AI in cybersecurity. The technical and operational challenges of using AI in cybersecurity were identified and discussed in Section 3. Section 4 presents the different ethical implications of AI in cybersecurity including privacy concerns, bias, accountability, and responsibility. Lessons learned from the reviews are presented in Section 5 including recommendations for further research directions. Finally, Section 6 concludes the paper.

2. Applications of AI in Cybersecurity

This section reviews the specific applications of AI in cybersecurity including intrusion detection systems (IDSs), malware analysis and prevention, threat intelligence and prediction, and user authentication and fraud detection, as illustrated in Figure 1.

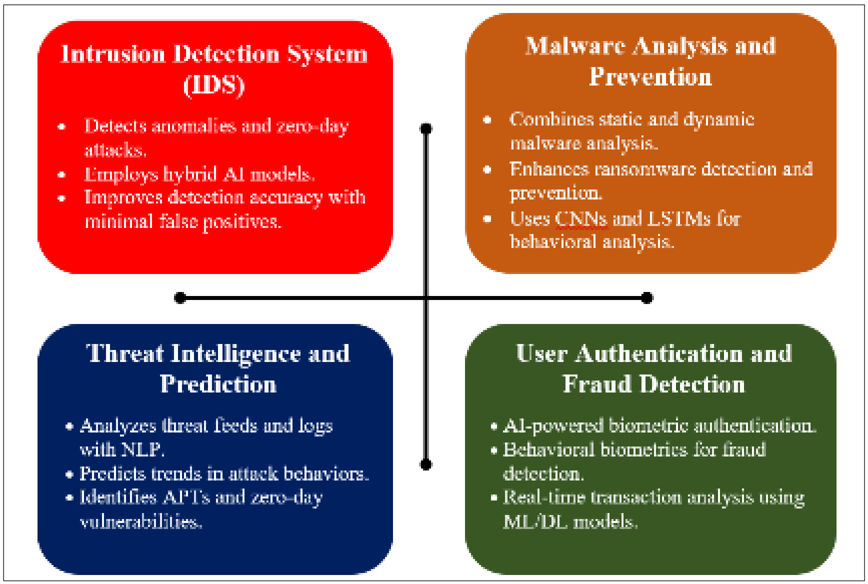
Intrusion Detection System (IDS)
• Detects anomalies and zero-day attacks.
• Employs hybrid AI models.
• Improves detection accuracy with minimal false positives.
Malware Analysis and Prevention
• Combines static and dynamic malware analysis.
• Enhances ransomware detection and prevention.
• Uses CNNs and LSTMs for behavioral analysis.
Threat Intelligence and Prediction
• Analyzes threat feeds and logs with NLP.
• Predicts trends in attack behaviors.
• Identifies APTs and zero-day vulnerabilities.
User Authentication and Fraud Detection
• AI-powered biometric authentication.
• Behavioral biometrics for fraud detection.
• Real-time transaction analysis using ML/DL models.

Figure 1 Key Applications of Artificial Intelligence in Cybersecurity

2.1. Intrusion Detection System (IDS)

Intrusion Detection Systems (IDS) are designed to monitor network traffic and detect suspicious activities or policy violations. Traditional IDS methods rely heavily on rule-based systems and static signatures, which are limited in detecting novel or evolving threats. AI, particularly ML and DL, has revolutionized IDS by enabling systems to identify anomalous patterns and adapt to emerging attack vectors. Numerous standalone and hybrid IDS models have been proposed in the literature to enhance anomaly detection, including zero-day attacks. For instance, an innovative hybrid model integrating Convolutional Neural Networks (CNNs), Gated Recurrent Units (GRUs), and a bi-directional Long Short-Term Memory (BiLSTM) network achieved 99.31% and 99.12% accuracy for multi-class and binary classifications, respectively, when evaluated using the NSL-KDD dataset [13]. Similarly, CNNs were combined with LSTM networks for spatial and sequential traffic analysis, achieving over 99% accuracy in detecting known and zero-day attacks on NSL-KDD and SDN-5-IoT benchmarks [14].

A comprehensive review of AI applications in IDS by [15] highlighted the significant contributions of AI in enhancing detection accuracy and response capabilities. Another systematic review of Network IDS (NIDS) solutions from 2016 to 2021 [16] analyzed various AI-based approaches, focusing on their methodologies, datasets, and evaluation metrics. This review observed a growing trend in hybrid and DL-based approaches, which have demonstrated superior performance compared to traditional methods. However, it also noted a critical limitation that many proposed solutions rely on outdated datasets, limiting their applicability to modern threat scenarios. While AI-driven IDS solutions offer notable advantages, such as scalability, adaptability, and improved detection accuracy, they also face significant challenges. These include the need for large, high-quality datasets for training, susceptibility to adversarial attacks, and the high computational costs associated with real-time detection. Addressing these challenges will be essential to fully harness the potential of AI in IDS.

2.2. Malware Analysis and Prevention

Malware analysis and prevention are essential components of cybersecurity, focusing on the identification, classification, and mitigation of malicious software that poses a threat to systems and networks. AI has significantly advanced both static and dynamic malware analysis techniques. In static analysis, features such as opcode sequences, API calls, and file metadata are extracted and analyzed without executing the malware [17]. Dynamic analysis, on the other hand, involves monitoring malware behavior in controlled environments such as sandboxes, where AI models such as Recurrent Neural Networks (RNNs) and LSTM networks have proven effective in capturing temporal dependencies in system call sequences [18]. By analyzing both static and dynamic features, hybrid AI models have emerged as robust solutions, improved detection accuracy, and reduced false positives. For instance, a CNN-LSTM model proposed in [19] achieved 99% accuracy for real-time malware detection, outperforming traditional approaches such as Support Vector Machines (SVM) and Decision Trees (DT). Similarly, a CNN-SVM model introduced in [20] demonstrated a high detection accuracy of 92.37% for malware threats.

The role of AI in malware detection with a focus on ransomware has also been extensively reviewed, with studies [21,22] emphasizing the importance of high-quality and authentic features in achieving reliable detection. These reviews highlighted that while AI models are powerful, their effectiveness is often limited by the quality of training datasets. Advanced feature engineering and the use of diverse, up-to-date datasets are critical for ensuring robust performance in real-world scenarios. Despite the significant progress, challenges such as high computational costs, the need for labeled datasets, and vulnerability to adversarial attacks persist.

2.3. Threat Intelligence and Prediction

Threat intelligence and prediction involve the collection, analysis, and interpretation of data to identify potential cybersecurity threats and predict future attack patterns. AI has significantly advanced this domain by automating threat detection, pattern recognition, and proactive defense mechanisms. By analyzing vast amounts of structured and unstructured data from sources such as threat feeds, logs, and dark web forums, AI uncovers hidden patterns and relationships that are critical for effective threat intelligence. For instance, AI models leveraging Natural Language Processing (NLP) can process textual data from threat reports and social media, extracting actionable insights to enhance cybersecurity preparedness [23]. Predictive analytics further augments this process by identifying trends in attack behaviors, enabling organizations to anticipate threats before they manifest.

AI-driven threat intelligence systems have demonstrated significant success in identifying advanced persistent threats (APTs) and zero-day vulnerabilities. For example, graph-based machine learning models have been employed to map relationships between attack indicators, providing early warnings about potential breaches [24]. DL models have also been utilized for accurate and timely detection of APT attacks. In [25], a DL model with automatic multi-layered feature extraction achieved 98.85% accuracy in detecting and classifying APT attacks on the NSL-KDD dataset, outperforming models such as the C5.0 decision tree and Bayesian network. Similarly, a Weighted PCA-based Enhanced Deep Neural Network (WPCA_E-DNN) model proposed in [26] achieved 95.2% accuracy in identifying APT characteristics using the CICAPT IIoT 2024 dataset. Furthermore, explainable DL models have addressed key challenges in traditional intrusion detection systems, such as low detection accuracy, high false-positive rates, and difficulties in identifying unknown or early-stage attacks, as reviewed in [27]. These advancements highlight the potential of integrating DL techniques with threat intelligence platforms to improve the accuracy of threat prediction, particularly for complex multi-stage attacks. However, challenges persist, including the need for high-quality data, the complexity of integrating AI systems into existing cybersecurity frameworks, and the risk of false positives or negatives.

2.4. User Authentication and Fraud Detection

User authentication and fraud detection are critical components of cybersecurity, focusing on verifying user identities and preventing unauthorized access or fraudulent activities. Traditional methods such as passwords and PINs are increasingly vulnerable to phishing, credential stuffing, and brute force attacks [28,29]. AI-powered biometric systems analyze physical and behavioral traits, such as fingerprints, facial features, voice patterns, and keystroke dynamics, for robust identity verification. For example, CNNs have achieved high accuracy in facial recognition by extracting intricate features from high-resolution images [30,31]. Similarly, behavioral biometrics based on ML monitor user interactions, such as typing speed, mouse movements, and touchscreen gestures, to detect anomalies indicative of fraud [32,33].

AI-based fraud detection systems have further enhanced security by analyzing transactional data in real time to identify suspicious patterns. RNNs and LSTM networks are widely used to model sequential transaction data, effectively identifying deviations from typical user behavior [34,35]. For instance, in [36], a Modified Binary Bat Algorithm (MBBA) was employed for feature selection, and Random Forest (RF) achieved the highest accuracy of 99.945% compared to Support Vector Machines (SVM) and Decision Trees (DT) for credit card fraud detection. Similarly, [37] proposed a hybrid DL model integrating CNNs, LSTM, and Transformers as base learners, with Extreme Gradient Boosting (XGBoost) as the meta-learner. This ensemble significantly outperformed individual base learners and traditional methods, achieving a sensitivity of 0.961, a specificity of 0.999, and an area under the receiver operating characteristic curve (AUC-ROC) of 0.972 using the European Credit Card Dataset.

3. Challenges of Using AI in Cybersecurity

The adoption of AI in cybersecurity has shown immense potential, but it is not without challenges. These challenges can be broadly categorized into technical and operational dimensions, each of which poses significant hurdles to the effective and efficient deployment of AI-based solutions.

3.1. Technical Challenges

These challenges encompass issues related to data, model performance, system scalability, and integration with existing infrastructure.

3.1.1. Data Availability and Quality

AI models rely heavily on high-quality, labeled datasets for training and evaluation. In cybersecurity, obtaining such datasets is particularly challenging due to the sensitive nature of the data and privacy concerns. Publicly available datasets, such as NSL-KDD and CICIDS2017, are often outdated and may not capture the dynamic nature of modern cyber threats. Additionally, these datasets are most imbalanced between benign and malicious samples, leading to biased models that perform poorly in real-world scenarios. To mitigate these issues, synthetic data generation techniques and federated learning have been proposed. While synthetic data can help address the imbalance problem, it lacks the authenticity of real-world data, potentially affecting model performance. Federated learning offers a privacy-preserving alternative by enabling collaborative training across decentralized datasets, but it introduces challenges such as communication overhead and the risk of model divergence.

3.1.2. Model Interpretability and Explainability

Many AI models, especially those based on deep learning, often lack transparency and interpretability which as known as "black boxes," making it difficult to understand their decision-making processes. This lack of interpretability hinders trust and adoption among cybersecurity professionals, who must justify and validate AI-driven decisions. Explainable AI (XAI) techniques, such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME), have been explored to enhance transparency, but they are computationally intensive and not always feasible for real-time applications. Furthermore, the trade-off between explainability and performance remains a critical issue, as highly interpretable models may not achieve the same level of accuracy as more complex architectures.

3.1.3. Adversarial Attacks on AI Models

Adversarial attacks, where malicious actors craft inputs designed to deceive AI models, represent a significant threat. For instance, slight perturbations in network traffic patterns or malware signatures can cause AI models to misclassify threats, leading to false negatives. Defensive strategies, such as adversarial training and robust model architectures, have shown promise but often come at the cost of increased computational complexity and reduced performance on benign inputs. Additionally, current countermeasures rely on pre-trained models or fixed thresholds that cannot adapt to new or evolving adversarial strategies necessitating continuous updates to defensive mechanisms, which can strain resources and complicate deployment. Emerging approaches, such as automated adversarial detection systems and the integration of ensemble defenses, are being developed to address these evolving threats, though they remain an area of active research.

3.1.4. Scalability and Deployment in Real-Time Systems

Cybersecurity systems must operate in real time to detect and mitigate threats promptly. However, the computational demands of AI models, particularly deep learning architectures, often pose significant obstacles to achieving seamless real-time deployment. Techniques such as model compression, pruning, quantization, and edge AI have been proposed to improve scalability and efficiency. While these methods can reduce computational overhead, they may inadvertently compromise model accuracy and robustness, creating trade-offs that must be carefully managed. Furthermore, scaling AI systems to handle high-throughput environments, such as large-scale enterprise networks or cloud infrastructures, presents a persistent challenge. Emerging solutions, including distributed AI frameworks, adaptive resource allocation, and hybrid edge-cloud architectures, are under active development to address these scalability constraints. These approaches aim to ensure that AI-driven cybersecurity solutions can maintain high performance while efficiently utilizing available resources.

3.1.5. Integration with Legacy Cybersecurity Systems

Legacy cybersecurity systems, which many organizations depend on, were not designed with modern AI technologies in mind. Achieving seamless compatibility and interoperability between these systems and AI-based solutions often demands extensive modifications, significant resource allocation, and expert intervention. Middleware solutions and API-based integrations are frequently employed to bridge these gaps, but they can introduce unintended latency and security vulnerabilities, compounding the challenges. Additionally, the absence of standardized protocols for integration further complicates the effective incorporation of AI into existing workflows. While international cybersecurity organizations are working on developing universal standards and frameworks, the slow pace of adoption and varying compliance levels across industries remain persistent barriers to progress.

3.2. Operational Challenges

These challenges often stem from resource limitations, workforce constraints, and the dynamic nature of cybersecurity threats. Addressing these challenges is crucial to ensure the sustained effectiveness and reliability of AI-based solutions.

3.2.1. High Computational Resource Requirements

AI models, especially those based on deep learning, require substantial computational resources for training and inference. This can be a significant barrier for small and medium-sized enterprises (SMEs) that lack access to high-performance computing infrastructure. While cloud-based AI solutions provide an alternative, they introduce challenges related to data security, latency, and potential compliance issues. Efficient utilization of computational resources without compromising performance is essential for the wider adoption of AI in cybersecurity. Innovations in hardware acceleration, such as GPUs and TPUs, have shown promise in addressing these computational demands. However, these technologies frequently entail high acquisition costs, specialized technical expertise, and maintenance overhead, making them less accessible to resource-constrained organizations. Emerging solutions, such as edge AI and low-power AI chips, aim to bridge this gap, but their integration into existing workflows remains an ongoing challenge.

3.2.2. False Positives and Negatives in Detection

One of the critical operational challenges is the high rate of false positives and negatives in AI-driven detection systems. False positives can overwhelm security teams with excessive alerts, leading to alert fatigue and diminished operational efficiency, while false negatives can result in undetected threats, potentially causing severe breaches. Continuous model retraining and the use of ensemble methods have been proposed to address this issue, but they require constant access to updated and diverse datasets alongside significant computational resources. Furthermore, the lack of standardized and universal evaluation metrics for detection systems complicates the benchmarking and comparison of model performance, hindering the ability to assess advancements effectively. Recent research has explored adaptive learning systems capable of dynamically adjusting detection thresholds based on evolving patterns in contextual data, demonstrating promise in improving both detection accuracy and operational resilience.

3.2.3. Lack of Standardized Frameworks for AI in Cybersecurity

The absence of a universally accepted framework for developing, deploying, and evaluating AI systems in cybersecurity poses significant challenges. This lack of standardization results in inconsistent performance, benchmarking difficulties, and regulatory compliance issues. Organizations often struggle to align their AI-driven cybersecurity measures with evolving legal and regulatory requirements, increasing the risk of legal and financial liabilities. Efforts to establish industry standards are underway. For instance, the National Institute of Standards and Technology (NIST) has proposed guidelines for trustworthy AI systems that prioritize fairness, transparency, and accountability. However, these standards are still in the early adoption phase. Developing clear, widely accepted guidelines for data handling, model evaluation, and deployment is critical to building trust and ensuring consistent, reliable AI-driven cybersecurity solutions.

3.2.4. Dynamic Cyber Threat Environment

AI systems in cybersecurity face a dynamic and ever-evolving threat environment where attackers increasingly leverage sophisticated techniques, including AI-driven tools, to bypass traditional defenses. This rapidly changing threat environment necessitates a constant need for AI models to be updated and retrained to remain effective. However, such updates can introduce significant operational overhead, consume resources, and sometimes destabilize existing systems. Additionally, the constant evolution of attack methods, such as adversarial attacks and polymorphic malware, further complicates the challenge. To address these issues, exploring and implementing continuous learning paradigms, such as online learning and incremental model updates, becomes crucial. These approaches enable AI systems to adapt in near real-time while maintaining robustness, ensuring that defenses evolve in parallel with emerging threats.

4. Ethical Implications of AI in Cybersecurity

Key ethical concerns revolve around the collection and use of sensitive data, transparency in decision-making, unintended biases, and the potential misuse of AI technologies. These issues require careful consideration to maintain trust, fairness, and security in digital systems. This section explores the critical ethical implications of deploying AI in cybersecurity.

4.1. Privacy Concerns

AI systems in cybersecurity often rely on large volumes of sensitive user data to train models effectively. This includes data such as network logs, communication metadata, and user behavior patterns. While this data is critical for improving model performance, its handling introduces privacy risks, including the potential for unauthorized access, data breaches, or misuse. For example, if sensitive information is not anonymized or adequately protected, there is a significant risk of exposing personally identifiable information (PII) or confidential corporate data. Furthermore, the deployment of AI systems for surveillance purposes, such as monitoring network traffic, user activities, or communication patterns, raises profound ethical questions about the balance between ensuring security and protecting individual privacy rights. These concerns are particularly critical in environments where users are unaware of the extent of monitoring or data collection. Implementing robust encryption protocols, applying differential privacy techniques to safeguard user data, and enforcing strict governance frameworks are crucial to ensure accountability and minimize risks of misuse.

4.2. Accountability and Responsibility

One of the most significant challenges with AI in cybersecurity is determining accountability when systems fail. For instance, if an AI system incorrectly flags legitimate activities as threats, false positives, or fails to detect actual threats, false negatives, the question arises who is responsible for the consequences? This issue is further complicated by the inherent opacity of many AI models, especially deep neural networks, which operate as "black boxes" due to their intricate and non-linear decision-making processes. This lack of transparency poses significant barriers to auditing and explaining AI-driven decisions, creating ethical problems related to trust, fairness, and the potential for errors. Addressing these concerns requires the establishment of clear accountability frameworks, encompassing comprehensive auditing tools, transparency mechanisms, and the adoption of XAI methodologies. Such measures are essential for fostering trust and ensuring that AI systems in cybersecurity remain reliable and equitable.

4.3. Bias and Discrimination

Bias in training data poses a significant ethical risk in AI-driven cybersecurity solutions. Models trained on biased datasets may produce unfair or inaccurate outcomes, such as disproportionately targeting certain groups or entities. For example, if historical data reflects discriminatory practices, the AI system may perpetuate these biases, leading to unequal treatment in cybersecurity measures. This could result in heightened scrutiny for certain demographics while neglecting others, thereby compromising the fairness and inclusivity of security protocols. Furthermore, biases can undermine the accuracy of threat detection mechanisms. If an AI model is skewed toward identifying threats based on specific patterns that align with biased data, it may fail to detect novel or less-represented attack vectors. This not only leaves certain entities more vulnerable to attacks but also erodes trust in AI-driven solutions. Addressing this issue requires proactive measures, including the adoption of fairness-aware algorithms that aim to minimize bias during training and prediction phases. Regular audits of training datasets and model outputs are also essential to identify and mitigate biases. Moreover, fostering diversity in the teams developing these systems can contribute to more equitable AI solutions by incorporating diverse perspectives into the design and implementation processes.

4.4. Dual-Use Problems

AI tools developed for cybersecurity can be repurposed for malicious activities, presenting a classic dual-use problem. For instance, AI algorithms designed to detect vulnerabilities could be reverse-engineered to exploit those same weaknesses. Similarly, offensive AI tools, such as automated attack frameworks and AI-driven malware, raise ethical concerns about their development and potential misuse. The dual-use nature of AI tools necessitates a multifaceted approach to ethical oversight. Developers must implement strict access controls to prevent unauthorized use of these technologies. Usage policies should clearly define the permissible applications of AI tools, with mechanisms to monitor and enforce compliance. International cooperation is also critical to address the global nature of cybersecurity threats. Establishing treaties and agreements that regulate the use and dissemination of dual-use AI technologies can help mitigate the risks of misuse. Additionally, embedding ethical considerations into the research and development process, such as conducting impact assessments, can guide the responsible creation and deployment of AI-driven cybersecurity tools.

Table 1 Summary of Challenges and Ethical Implications of Using AI in Cybersecurity

Categories	Challenges	Implications	Mitigation Strategy	Associated Risks
Technical Challenges	Data Availability and Quality	Outdated or imbalanced datasets lead to poor model performance.	Use synthetic data generation and federated learning.	Synthetic data lacks authenticity; federated learning introduces communication overhead and model divergence.
	Model Interpretability and Explainability	Lack of transparency reduces trust and adoption.	Implement XAI techniques such as SHAP and LIME.	XAI techniques are computationally intensive and may not support real-time applications.
	Adversarial Attacks	Malicious actors can deceive AI models, resulting in false negatives.	Employ adversarial training and ensemble models.	Increased computational cost and potential reduction in model generalizability.
	Scalability and Real-Time Performance	High computational costs limit scalability in real-time applications.	Use model compression, pruning, and edge AI.	Compression may compromise accuracy; edge AI requires specialized hardware.
	Integration with Legacy Systems	Difficulty in merging AI with outdated systems increases costs and complexity.	Develop standardized APIs and middleware solutions.	Middleware can introduce latency and potential security vulnerabilities.
Operational Challenges	High Computational Requirements	Small organizations struggle to deploy AI due to resource constraints.	Utilize cloud-based solutions, hardware accelerators, and low-power AI chips.	Cloud solutions raise data security concerns; accelerators are costly and require expertise.
	False Positives and Negatives	Excessive alerts cause fatigue, while	Adopt adaptive learning systems and ensemble methods.	Requires constant access to updated datasets and

		undetected threats lead to breaches.		significant computational resources.
	Lack of Standardized Frameworks	Inconsistent performance and benchmarking difficulties arise.	Promote industry-wide standards such as NIST’s trustworthy AI guidelines.	Adoption of standards may be slow across industries.
Ethical Implications	Privacy Concerns	Collection of sensitive data risks breaches and unauthorized access.	Implement encryption, differential privacy, and strict governance frameworks.	Privacy techniques may reduce data utility for training models.
	Accountability and Responsibility	Lack of clear accountability when AI systems fail or misclassify.	Establish accountability frameworks and use XAI to make decisions auditable.	Defining responsibility in multi-stakeholder systems is challenging.
	Bias and Discrimination	Biased datasets result in unfair outcomes and discriminatory practices.	Conduct regular audits and adopt fairness-aware algorithms, and ensure diverse development teams.	Auditing processes can be resource-intensive; fairness-aware algorithms may impact model performance.
	Dual-Use Problems	AI tools can be misused for malicious purposes.	Enforce strict access controls, and usage policies, and establish international treaties.	Policies may lag behind technological advancements; enforcement is resource-intensive.

5. Further Research Directions

This section identifies and discusses recommended further research directions addressing the challenges and ethical implications of the integration of AI in cybersecurity.

5.1. Developing Diverse and Up-to-Date Datasets

As mentioned earlier, the existing cybersecurity datasets used to develop AI models are outdated, and imbalanced, and often fail to capture the complexity and dynamism of real-world attack vectors, leading to models with limited generalizability. To address this, it is essential to incorporate data representing polymorphic malware, APTs, and zero-day exploits. Therefore, future research on AI applications in cybersecurity should prioritize creating diverse and up-to-date datasets tailored to the dynamic nature of cyber threats. Equally important is the exploration of privacy-preserving techniques, such as FL and differential privacy, to enable organizations to share threat intelligence without risking sensitive information. Additionally, synthetic data generation methods require further refinement to replicate the behavior of sophisticated cyberattacks, including fileless malware and multi-stage attacks, ensuring models are trained on realistic scenarios.

5.2. Enhancing Explainability in AI Models for Cybersecurity

Improving the explainability of AI models in cybersecurity is a critical area of research. Advanced models like Deep Neural Networks (DNNs) often deliver exceptional performance but are hindered by their "black-box" nature, which limits trust and operational adoption in sensitive contexts. Future efforts should prioritize developing interpretable AI techniques, such as SHAP, LIME, and counterfactual explanations, tailored to the unique challenges of cybersecurity. These techniques must provide contextualized, actionable insights that help practitioners identify root causes and implement effective mitigation strategies. Additionally, optimizing these methods to reduce computational overhead is essential to ensure their practicality for real-time applications.

5.3. Resilience Against Adversarial Attacks

Resilience against adversarial attacks represents a growing concern in AI applications for cybersecurity. Models are increasingly susceptible to adversarial manipulations, such as evasion attacks or data poisoning. Research must prioritize the development of robust training methods, including adversarial training and gradient masking, to enhance model resistance to such threats. Dynamic defense mechanisms capable of detecting and mitigating adversarial behaviors in real time are also crucial. To support this, establishing comprehensive evaluation frameworks will allow for consistent benchmarking of AI model resilience against various adversarial conditions.

5.4. Developing Frameworks to Govern the Ethical Use of AI in Cybersecurity

The integration of AI into cybersecurity offers significant advancements but raises ethical concerns, including privacy violations, algorithmic biases, and accountability for automated decisions. Developing ethical frameworks is essential to address these challenges by ensuring transparency, accountability, and fairness in AI-driven systems. Such frameworks should promote interpretable decision-making, hold developers and organizations responsible for the consequences of AI deployment, and implement measures to detect and mitigate biases that could lead to discriminatory outcomes. Future research should focus on the development of these ethical governance frameworks for AI usage in cybersecurity. These frameworks can balance innovation with societal values such as privacy, security, and fairness, ensuring responsible and trustworthy use of AI in cybersecurity.

5.5. Real-Time and Edge AI for Cybersecurity

The rise of IoT devices and edge computing necessitates real-time AI solutions that operate with minimal latency and resource constraints. Designing lightweight AI models optimized for low-power devices is a critical area of research, enabling on-device threat detection without over-reliance on centralized infrastructure. Decentralized architectures, such as blockchain-based collaborative detection networks, can further enhance scalability and resilience. Achieving these goals will require innovations in latency optimization, such as model compression techniques such as pruning and quantization, alongside faster inference engines tailored to cybersecurity applications.

6. Conclusion

AI has transformed the way networks and digital systems are protected by providing enhanced threat detection, predictive capabilities, and automated responses. While these advancements demonstrate the potential of AI to address complex cyber threats, they also introduce significant challenges. This paper investigated the technical, operational, and ethical issues such as data quality, model interpretability, adversarial attacks, and the dynamic cyber threat environment highlighting the need for robust, scalable, and transparent AI solutions. From an ethical perspective, concerns around privacy, accountability, bias, and dual-use remain critical. Addressing these challenges requires a collaborative effort among researchers, policymakers, and industry practitioners to establish standardized frameworks and guidelines for the responsible development and deployment of AI in cybersecurity. Future research should focus on creating diverse, up-to-date datasets, improving model explainability, and enhancing resilience against adversarial attacks. Moreover, exploring lightweight, real-time AI solutions tailored for edge computing and IoT environments will be essential as the digital system continues to expand.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Al Kuwaiti, A.; Nazer, K.; Al-Reedy, A.; Al-Shehri, S.; Al-Muhanna, A.; Subbarayalu, A.V.; Al Muhanna, D.; Al-Muhanna, F.A. A Review of the Role of Artificial Intelligence in Healthcare. J. Pers. Med. 2023, 13, 951, doi:10.3390/jpm13060951.

[2] Alozie, E.; Olagunju, H.I.; Faruk, N.; Garba, S. Technical Considerations of Federated Learning in Digital Healthcare Systems. In Federated Learning for Digital Healthcare Systems; Elsevier, 2024; pp. 237–282 ISBN 978-0-443-13897-3.

[3] Plathottam, S.J.; Rzonca, A.; Lakhnori, R.; Iloeje, C.O. A Review of Artificial Intelligence Applications in Manufacturing Operations. J. Adv. Manuf. Process. 2023, 5, e10159, doi:10.1002/amp2.10159.

[4] Chiu, T.K.F.; Xia, Q.; Zhou, X.; Chai, C.S.; Cheng, M. Systematic Literature Review on Opportunities, Challenges, and Future Research Recommendations of Artificial Intelligence in Education. Comput. Educ. Artif. Intell. 2023, 4, 100118, doi:10.1016/j.caeai.2022.100118.

[5] Abdulkarim, A.; Faruk, N.; Alozie, E.; Sowande, Olugbenga.A.; Olayinka, I.-F.Y.; Usman, A.D.; Adewole, K.S.; Oloyede, A.A.; Chiroma, H.; Garba, S.; et al. Application of Machine Learning Algorithms to Path Loss Modeling: A Review. In Proceedings of the 2022 5th Information Technology for Education and Development (ITED); IEEE: Abuja, Nigeria, November 1 2022; pp. 1–6.

[6] Chiroma, H.; Nickolas, P.; Faruk, N.; Alozie, E.; Olayinka, I.-F.Y.; Adewole, K.S.; Abdulkarim, A.; Oloyede, A.A.; Sowande, O.A.; Garba, S.; et al. Large Scale Survey for Radio Propagation in Developing Machine Learning Model for Path Losses in Communication Systems. Sci. Afr. 2023, 19, e01550, doi:10.1016/j.sciaf.2023.e01550.

[7] Javaid, M.; Haleem, A.; Khan, I.H.; Suman, R. Understanding the Potential Applications of Artificial Intelligence in Agriculture Sector. Adv. Agrochem 2023, 2, 15–30, doi:10.1016/j.aac.2022.10.001.

[8] Pandey, U.; Pathak, A.; Kumar, A.; Mondal, S. Applications of Artificial Intelligence in Power System Operation, Control and Planning: A Review. Clean Energy 2023, 7, 1199–1218, doi:10.1093/ce/zkad061.

[9] Weber, P.; Carl, K.V.; Hinz, O. Applications of Explainable Artificial Intelligence in Finance—a Systematic Review of Finance, Information Systems, and Computer Science Literature. Manag. Rev. Q. 2024, 74, 867–907, doi:10.1007/s11301-023-00320-0.

[10] Straub, V.J.; Morgan, D.; Bright, J.; Margetts, H. Artificial Intelligence in Government: Concepts, Standards, and a Unified Framework. Gov. Inf. Q. 2023, 40, 101881, doi:10.1016/j.giq.2023.101881.

[11] Banerjee, A.; Kabadi, S.; Karimov, D. The Transformative Power of AI: Projected Impacts on the Global Economy by 2030. Rev. Artif. Intell. Educ. 2023, 4, e020, doi:10.37497/rev.artif.intell.educ.v4i00.20.

[12] Olagunju, H.I.; Alozie, E.; Imoize, A.L.; Faruk, N.; Garba, S.; Baba, B.A. Cybersecurity in the Internet of Medical Things for Healthcare Applications. In Cybersecurity in Emerging Healthcare Systems; Imoize, A.L., Meshram, C., Awotunde, J.B., Farhaoui, Y., Do, D.-T., Eds.; Institution of Engineering and Technology, 2024; pp. 41–74 ISBN 978-1-83953-951-0.

[13] Zulfiqar, Z.; Malik, S.U.R.; Moqurrab, S.A.; Zulfiqar, Z.; Yaseen, U.; Srivastava, G. DeepDetect: An Innovative Hybrid Deep Learning Framework for Anomaly Detection in IoT Networks. J. Comput. Sci. 2024, 83, 102426, doi:10.1016/j.jocs.2024.102426.

[14] Ali, R.M.; Ram Baheti, M. Enhancing IoT Security: A Study on Hybrid Intrusion Detection Methods. In Proceedings of the 2024 IEEE 3rd World Conference on Applied Intelligence and Computing (AIC); IEEE: Gwalior, India, July 27 2024; pp. 1373–1380.

[15] Markevych, M.; Dawson, M. A Review of Enhancing Intrusion Detection Systems for Cybersecurity Using Artificial Intelligence (AI). Int. Conf. Knowl.-BASED Organ. 2023, 29, 30–37, doi:10.2478/kbo-2023-0072.

[16] Habeeb, M.S.; Babu, T.R. Network Intrusion Detection System: A Survey on Artificial Intelligence-based Techniques. Expert Syst. 2022, 39, e13066, doi:10.1111/exsy.13066.

[17] Pan, Y.; Ge, X.; Fang, C.; Fan, Y. A Systematic Literature Review of Android Malware Detection Using Static Analysis. IEEE Access 2020, 8, 116363–116379, doi:10.1109/ACCESS.2020.3002842.

[18] Redhu, A.; Choudhary, P.; Srinivasan, K.; Das, T.K. Deep Learning-Powered Malware Detection in Cyberspace: A Contemporary Review. Front. Phys. 2024, 12, 1349463, doi:10.3389/fphy.2024.1349463.

[19] Akhtar, M.S.; Feng, T. Detection of Malware by Deep Learning as CNN-LSTM Machine Learning Techniques in Real Time. Symmetry 2022, 14, 2308, doi:10.3390/sym14112308.

[20] Shoniwa, M.; Veerabudren, K.; Sharma, M. AI-Based Malware Threat Prediction through CNN-SVM Ensemble. In Proceedings of the 2024 International Conference on Next Generation Computing Applications (NextComp); IEEE: Mauritius, October 24 2024; pp. 1–6.

[21] Ferdous, J.; Islam, R.; Mahboubi, A.; Zahidul Islam, M. AI-Based Ransomware Detection: A Comprehensive Review. IEEE Access 2024, 12, 136666–136695, doi:10.1109/ACCESS.2024.3461965.

[22] Gaber, M.G.; Ahmed, M.; Janicke, H. Malware Detection with Artificial Intelligence: A Systematic Literature Review. ACM Comput. Surv. 2024, 56, 1–33, doi:10.1145/3638552.

[23] Natural Language Processing (NLP) for Threat Intelligence. In Advances in Information Security, Privacy, and Ethics; IGI Global, 2024; pp. 247–262 ISBN 979-8-3693-9491-5.

[24] Alwasel, B.; Aldribi, A.; Alreshoodi, M.; Alsukayti, I.S.; Alsuhaibani, M. Leveraging Graph-Based Representations to Enhance Machine Learning Performance in IIoT Network Security and Attack Detection. Appl. Sci. 2023, 13, 7774, doi:10.3390/app13137774.

[25] Hassannataj Joloudari, J.; Haderbadi, M.; Mashmool, A.; Ghasemigol, M.; Band, S.S.; Mosavi, A. Early Detection of the Advanced Persistent Threat Attack Using Performance Analysis of Deep Learning. IEEE Access 2020, 8, 186125–186137, doi:10.1109/access.2020.3029202.

[26] Enhanced Deep Learning for IIoT Threat Intelligence: Revealing Advanced Persistent Threat Attack Patterns. In Communications in Computer and Information Science; Springer Nature Singapore: Singapore, 2025; pp. 201–217 ISBN 978-981-97-9742-4.

[27] Mutalib, N.H.A.; Sabri, A.Q.M.; Wahab, A.W.A.; Abdullah, E.R.M.F.; AlDahoul, N. Explainable Deep Learning Approach for Advanced Persistent Threats (APTs) Detection in Cybersecurity: A Review. Artif. Intell. Rev. 2024, 57, doi:10.1007/s10462-024-10890-4.

[28] Adeniran, T.C.; Jimoh, R.G.; Abah, E.U.; Faruk, N.; Alozie, E.; Imoize, A.L. Vulnerability Assessment Studies of Existing Knowledge-Based Authentication Systems: A Systematic Review. Sule Lamido Univ. J. Sci. Technol. 2024, 8, 34–61, doi:10.56471/slujst.v7i.485.

[29] Adeniran, T.C.; Jimoh, R.G.; Abah, E.U.; Faruk, N.; Alozie, E. Evaluation of an Enhanced Dynamic Knowledge-Based Authentication System (EDKBA) in the Era of Social Media. In Proceedings of the 2023 2nd International Conference on Multidisciplinary Engineering and Applied Science (ICMEAS); IEEE: Abuja, Nigeria, November 1 2023; pp. 1–5.

[30] Ali, W.; Tian, W.; Din, S.U.; Iradukunda, D.; Khan, A.A. Classical and Modern Face Recognition Approaches: A Complete Review. Multimed. Tools Appl. 2021, 80, 4825–4880, doi:10.1007/s11042-020-09850-1.

[31] Almabdy, S.; Elrefaei, L. Deep Convolutional Neural Network-Based Approaches for Face Recognition. Appl. Sci. 2019, 9, 4397, doi:10.3390/app9204397.

[32] Awad, A.I.; Babu, A.; Barka, E.; Shuaib, K. AI-Powered Biometrics for Internet of Things Security: A Review and Future Vision. J. Inf. Secur. Appl. 2024, 82, 103748, doi:10.1016/j.jisa.2024.103748.

[33] Bansal, P.; Ouda, A. Continuous Authentication in the Digital Age: An Analysis of Reinforcement Learning and Behavioral Biometrics. Computers 2024, 13, 103, doi:10.3390/computers13040103.

[34] Wiese, B.; Omlin, C. Credit Card Transactions, Fraud Detection, and Machine Learning: Modelling Time with LSTM Recurrent Neural Networks. In Innovations in Neural Information Paradigms and Applications; Bianchini, M., Maggini, M., Scarselli, F., Jain, L.C., Eds.; Studies in Computational Intelligence; Springer Berlin Heidelberg: Berlin, Heidelberg, 2009; Vol. 247, pp. 231–268 ISBN 978-3-642-04002-3.

[35] Benchaji, I.; Douzi, S.; Ouahidi, B.E. Credit Card Fraud Detection Model Based on LSTM Recurrent Neural Networks. J. Adv. Inf. Technol. 2021, 12, 113–118, doi:10.12720/jait.12.2.113-118.

[36] Olasupo, Y.A.; Malgwi, M.Y.; Hambali, M.A. Enhancing Credit Card Fraud Detection with Modified Binary Bat Algorithm: A Comparative Study with SVM, RF, and DT. J. Comput. Sci. Eng. JCSE 2024, 5, 80–98, doi:10.36596/jcse.v5i2.771.

[37] Ileberi, E.; Sun, Y. A Hybrid Deep Learning Ensemble Model for Credit Card Fraud Detection. IEEE Access 2024, 12, 175829–175838, doi:10.1109/ACCESS.2024.3502542.