technologies

MDPI

Review

Innovations and Future Perspectives in the Use of Artificial Intelligence for Cybersecurity: A Scoping Review

Cristian Randieri [1,2,*], Francesca Fiani [2], Kevin Lubrano [1] and Christian Napoli [2,3,4]

[1] Department of Theoretical and Applied Sciences, eCampus University, Via Isimbardi 10, 22060 Novedrate, Italy
[2] Department of Computer, Control and Management Engineering, Sapienza University of Rome, Via Ariosto 25, 00185 Rome, Italy; francesca.fiani@uniroma1.it (F.F.); cnapoli@diag.uniroma1.it (C.N.)
[3] Institute for Systems Analysis and Computer Science, Italian National Research Council, 00185 Rome, Italy
[4] Department of Artificial Intelligence, Czestochowa University of Technology, 42201 Częstochowa, Poland
* Correspondence: cristian.randieri@uniecampus.it

Abstract

Cybersecurity is a field in which integration of artificial intelligence (AI) represents a significant direction towards protection against cyber threats. This scoping review explores the current impact and future prospects of AI in four key areas of cybersecurity: threat detection, endpoint security, phishing and fraud detection, and network security. The main goal was to answer the research question, 'Is AI an effective method to enhance current infrastructures' cybersecurity?' **Method**: Through the PRISMA-ScR protocol, 2548 records were identified from the Google Scholar database from January 2020 to April 2025. The following search terms were used to identify available literature: "Artificial Intelligence Cybersecurity", "Machine Learning Cybersecurity", "Cybersecurity Innovation AI", "AI Future Perspective Cybersecurity", "Machine Learning Innovation Cybersecurity". The search only included articles in English. No grey literature has been included. Articles with a focus on performance optimization, cost analysis and business models without a focus on privacy and security have been discarded. **Results**: The impact and performance of AI algorithms have been highlighted through a selection of 20 articles. Both Machine Learning and Neural Network methods have been employed in the literature, with Decision Trees and Random Forest being the most common approaches. **Discussion**: The main common limitations of the analyzed articles have been discussed, highlighting possible future directions of research to tackle them. **Conclusions**: Despite the evidenced limitations, AI showed promising results in improving cybersecurity, especially concerning cyber attack detection and classification, with methods able to grant very high accuracy and trustworthiness.

Keywords: scoping review; PRISMA; cybersecurity; cyber attacks; artificial intelligence; machine learning

Academic Editor: Arash Habibi Lashkari

Received: 10 October 2025
Revised: 21 November 2025
Accepted: 7 December 2025
Published: 11 December 2025

Citation: Randieri, C.; Fiani, F.; Lubrano, K.; Napoli, C. Innovations and Future Perspectives in the Use of Artificial Intelligence for Cybersecurity: A Scoping Review. *Technologies* **2025**, *13*, 584. https://doi.org/10.3390/technologies13120584

1. Introduction

Cybersecurity is a field in computer science which has seen a drastic increase in attention in recent years, composed of a set of technologies, practices, and policies aimed at protecting information systems, networks, applications, and data from cyber attacks and unauthorized access [1]. Its main aim is to ensure the confidentiality, integrity, and availability of information, while at the same time preventing threats such as ransomware, malware, phishing, and data theft.

Cybersecurity is crucial in various sectors, including industry [2], healthcare [3], finance [4] and critical infrastructure [5]. In particular, Industry 4.0 represents a new era of industrial production, characterized by the integration of advanced technologies such as the Internet of Things (IoT), physical systems, and blockchain in the classic pipeline [6]. These interconnected systems improve efficiency and productivity, but at the same time increase the likelihood and attack space for cybercriminals. For this reason, cybersecurity is a field in which further studies are particularly relevant.

Blockchain, a decentralized digital ledger that stores data in a secure, transparent, and immutable way, offers innovative solutions for data security [7]. Each "block" of data is linked to the previous block, forming a chain, and is used to trace the provenance of products, improving transparency in supply chains. Smart grids, on the other hand, are an advanced electrical network that uses bidirectional communication and distributed intelligent devices to improve the efficiency and reliability of energy distribution by enabling real-time monitoring and remote control of electrical devices [8]. Both fields have shown promise, but can easily be subject to cyber attacks. Given the rise of use of Artificial Intelligence (AI) in several fields, it is particularly important to understand if such technologies can also benefit cybersecurity by further increasing protection against threats.

The aim of this research is therefore to explore the application of AI in cybersecurity to improve the protection of systems in these advanced fields, and therefore to answer to the question, 'Is AI an effective method to enhance current infrastructures' cybersecurity?' As this review will demonstrate, AI and machine learning (ML) can indeed offer advanced tools to identify, prevent and respond to threats more efficiently. For example, ML algorithms can detect anomalies and malicious activities, while AI can automate responses to incidents and attacks.

This scoping review differs from previous works in its specific focus on the integration of AI in cybersecurity for emerging sectors such as Industry 4.0, physical systems, IoT, Blockchain, and Smart Grids. While previous studies have examined the use of AI in cybersecurity in more general terms [9–11], this research focuses instead on how these technologies can be applied to address the practical challenges of these advanced and increasingly emerging fields, contextualizing the proposed AI solutions to real-world applications. Compared to other studies with a more domain-specific focus [12], our work instead offers a more complete cross-domain synthesis by integrating not only different domains (industry and infrastructure), but also on several attack types such as Man in the Middle, Denial of Service and malware, providing a complete overview of the shared commonalities and limitations of AI in cybersecurity. Moreover, a focus on the last five years (January 2020–April 2025) allowed us to vastly differ from older works [13] and to more easily identify the current gaps in research. Through a Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) methodology, 20 articles have been selected for a comparative analysis, showcasing how the integration of AI technology in cybersecurity is performed and highlighting current limitations of the literature. Several reviews in the literature on the topic do not make use of the PRISMA methodology to conduct the literature research [13–15], making our study not only more solid by reducing selection bias, but also reproducible. When selecting articles, our main focus was on high-impact (Q1/Q2) English publications in order to guarantee the relevance of the selected articles in the literature. Articles were selected through the Google Scholar database with specific keywords reported in the Methods section.

Various ML and deep learning algorithms have been examined, such as Support Vector Machines (SVM), Random Forest (RF), Decision Trees (DT), K-Nearest Neighbors (KNN), and deep neural networks such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), Autoencoders (AEs), as well

as hybrid approaches integrated with fuzzy logic. Recent studies have also demonstrated the applicability of lightweight CNNs in critical domains such as medical diagnostics [16]. These models have been analyzed in relation to their applicability in different fields, ranging from intrusion detection and zero-day attacks, to endpoint protection and phishing and fraud detection. The reviewed studies have used classical benchmarks, such as NSL-KDD, KDD Cup 99, UNSW-IDS15, and CICIDS2017, and datasets from real-world contexts, such as Supervisory Control And Data Acquisition (SCADA) systems, IoT networks, and smart grids, to validate the performance of the models. The overall results show promising performances, with accuracy rates often higher than 95–99% in the training and validation phase. However, a series of critical issues have also emerged, leading to the formulation of future lines of research, oriented towards the adoption of Explainable AI (xAI) techniques [17], the use of computationally light solutions (for example, through federated or few-shot learning) and the development of datasets more representative of the real dynamics of cyberspace.

The review will be structured as follows. Section 2 will analyze the method employed to retrieve the literature used in this scoping review. Section 3 will focus on the analysis of each work, providing a summary of the proposed methodologies, the employed datasets, the obtained results, and the limitations of each work. Section 4 will present a comprehensive analysis of the synthesized works, highlighting common traits, limitations both of literature and of the study and presenting possible future directions for research, concurrently answering the proposed research question. Finally, Section 5 will provide a summary of this scoping review, including some final considerations.

2. Methods

This review was performed following the PRISMA-ScR guidelines and checklist last updated in September 2019. This format of review was chosen in order to perform a precise analysis of current literature and identify gaps in the current state of the art, while also verifying if AI does indeed benefit cybersecurity performances and efficacy.

The research of articles and bibliography was carried out using Google Scholar Google LLC (Alphabet Inc., Mountain View, CA, USA), with the following search keys used in the research of relevant literature: “Artificial Intelligence Cybersecurity”, “Machine Learning Cybersecurity”, “Cybersecurity Innovation AI”, “AI Future Perspective Cybersecurity”, “Machine Learning Innovation Cybersecurity”. The research mainly focused on the methodologies adopted by AI for the prevention and protection against potential attacks such as Man-in-the-Middle (MitM), Intrusion Detection and Prevention, or False Data Injection. The research time frame is based on the last five years, from 2020 to 2025. All articles were searched exclusively in English. No grey literature has been considered in the research.

From the initial search carried out with Google Scholar and the keywords reported previously, 2548 potential articles and studies on the topics of AI and ML applied to Cybersecurity were identified and served as a basis for our investigation process. The last database search to update the potential article list was performed in April 2025.

The first step was to perform an analysis and elimination of 90 duplicate studies; this procedure began with the extraction of the list of titles with attached authors, bibliographies and year of publication through an application called Harzing Publish or Perish (Harzing B.V., Schagen, The Netherlands, version 8). Thanks to this software, it was possible to replicate the searches performed with the Google Scholar search engine and easily extract the complete bibliography to perform the subsequent duplicate elimination. The results in terms of numbers and titles were verified to match the direct search with Scholar. The bibliography from each search was then imported into the Zotero application (Corporation for Digital Scholarship (CDS), Vienna, VA, USA, version 7.0.13), which automatically

calculated the number of duplicates in the library and reported the specific duplicate files (same title or abstract) to verify the correctness of the resulting exclusion.

The second step was to select the articles and studies related only to the areas of Industry 4.0 and healthcare sectors, IoT, Blockchain, Smart Grids, Cyber Physical Systems and Wireless Sensor Networks systems, and Deep Learning, Explainable AI and Generative AI algorithms. Articles that deal with the performance of AI and ML algorithms for Cybersecurity by comparing their performance were not discarded. Similarly, methodologies adopted by AI for the prevention and protection against potential attacks, such as MitM, intrusion detection/prevention, or false data injection, were also accepted. Explainable AI has also been included in our review process due to the importance of transparency and explainability in cybersecurity. Transparency is fundamental to let operators understand why an AI system made a certain decision and evaluate its reliability and appropriateness, particularly in the medical context, while explainability can also help identify vulnerabilities and biases, possibly highlighting areas where to strengthen AI algorithms to protect them from cyber attacks. Explainable AI can also help address ethical and legal concerns such as data protection and fairness. Finance and Autonomous Vehicles, while highly differing from the industrial and infrastructure sectors due to strict regulations for the first (e.g., PCI DSS, GDPR) and safety standards for the second (e.g., ISO 26262, UNECE WP.29), have received much attention in recent years [18,19], leading us to also include them in our research despite the incompatibilities caused by the different framework. This allowed us to perform a more comprehensive overview of the state of the art, allowing us to identify common literature gaps considering all available AI for cybersecurity applications. Articles which focused on performance optimization, cost analysis and without a focus on data security and privacy were instead discarded, as we specifically aimed at including articles which delved into methods for data protection and attack identification. Preprints, journal and conference papers were included, while patents were excluded. The potential articles for the research total 43, having discarded 2419.

Finally, from the last 43 articles, which were all accessible, a further 19 studies were discarded for the following reasons:

- Arguments not treated (N = 5): Five articles do not delve into the themes of using Cybersecurity with AI in the areas described above.
- Survey not relevant (N = 4): Four articles are surveys on topics which do not belong to the ones identified for this scoping review.
- Low impact of publication (N = 10): Ten articles have a low publication quartile (Q3/Q4). To verify the quartile, a search was performed using Scimago (SCIMAGO RESEARCH GROUP, S.L., Granada, Spain), which returned, in the case of the presence of the publication journal, a value on a scale from Q1 to Q4. Q3 and Q4 results correspond to poor quality of the publication journal and a low impact in the literature of the published article due to a lower citation score, and have therefore been discarded as they are not significantly impactful to the current state of the art compared to Q1 and Q2 publications. However, it must be noted that a low quartile does not necessarily indicate low quality of the publication, but only their reduced impact in the literature.

One reviewer (K.L.) performed the screening and assessment of the articles, with a second reviewer (F.F.) verifying the process and confirming the selection of articles. In case of disagreements, the reviewers discussed the inclusion and/or exclusion of a certain article until they reached a consensus. In case of lack of an agreement, a third reviewer (C.R.) acted as an adjudicator. After the whole screening process, 24 articles remained for analysis and evaluation. Each article was narratively summarized, providing a report for each article of employed methodologies, datasets, numerical results (where applicable), and limitations.

Figure 1 illustrates the workflow of study selection of this scoping review. A summary of the inclusion and exclusion criteria for the selection of works has been presented in Table 1.

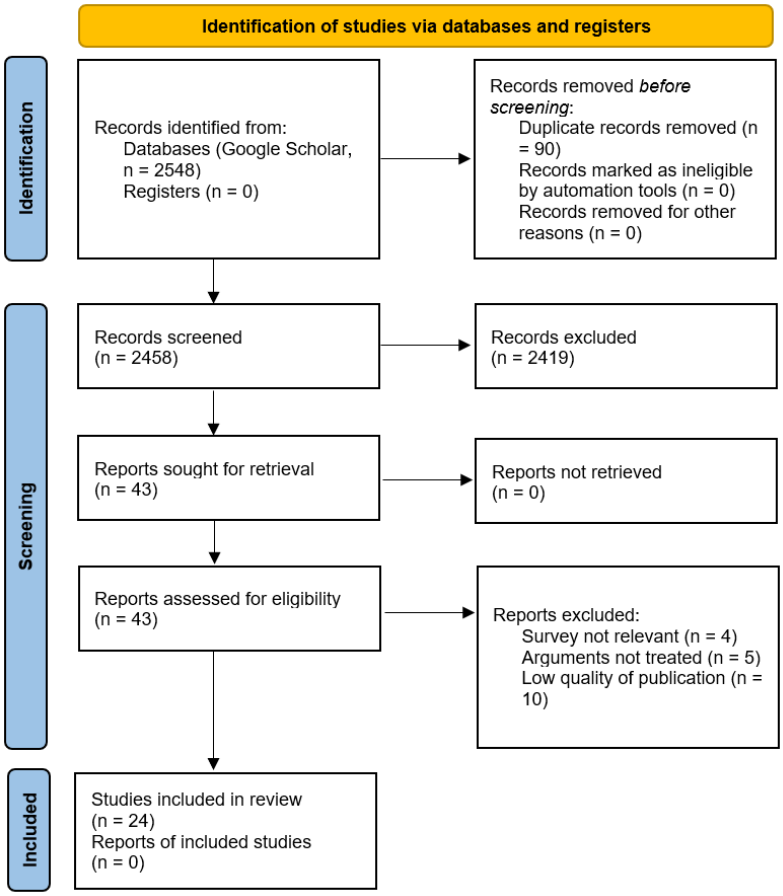

Figure 1. PRISMA -ScR diagram of the article selection process.

Table 1. Inclusion and exclusion criteria for the article selection through the PRISMA-ScR guidelines.

Inclusion Criteria	Exclusion Criteria
Preprint, journal or conference article	Patent, opinion, commentary
Work published in Q1/Q2 journal or in conference of rank $\geq$ B	Work published in Q3/Q4/unranked journal or in conference of rank < B/unranked
Work published in English	Work published in other languages
Work published $\geq$ 5 years (2020–2025)	Work published < 5 years (before 2020)
Articles in the following areas: IoT, Blockchain, Industry 4.0, Smart Grids, Wireless Sensor Network, Cyber Physical Systems, Healthcare, Finance, Autonomous Vehicles	Articles with a focus on: Performance Optimization, Cost Analysis, Models without security/privacy focus
Works on AI, ML or Deep Learning solutions for cybersecurity and/or for prevention and protection against cyber attacks (both comparative reviews and novel approaches)	Works which do not delve on AI for cybersecurity

3. Results

After screening, 24 works were selected for inclusion in the scoping review. Of them, 15 report numerical results (62.5%), with the remaining 9 (37.5%) presenting only theoretical analysis or reviews of previous works.

Analyzing more in detail the approaches used in the works which report numerical results, ten present ML approaches (66.67%, 41.67% of total works), while eleven use Deep Learning (DL) instead (73.33%, 45.83% of total works), with seven presenting both types of approaches in comparative analysis (46.67%, 29.17% of total works). Among all mentioned algorithms, DT and RF are the most frequently used in seven works each (46.67%, 29.17% of total works), followed by RNN in six works (40%, 25% of total works), AE, KNN, SVM and CNN in five works (33.33%, 20.83% of total works), Naive Bayes and Deep Belief Networks in four works (26.67%, 16.67% of total works), and Multilayer Perceptron, Deep Neural Networks, Isolation Forest and Linear Regression in two works (13.33%, 8.33% of total works). Approaches such as Principal Component Analysis and Extra Tree Classifier only appear in one work (6.6%, 4.17% of total works). Datasets tend to intersect scarcely, with NSL-KDD being used in five works (33.33%, 20.83% of total works), KDD Cup 99 in three (20%, 12.5% of total works), and UNSW-IDS15 in two (13.33%, 8.33% of total works).

Of the 24 works, only two deal with Industry 4.0 (8.33%), with another two mentioning industry in general (8.33%). Ten articles mention uses for IoT systems (41.67%), two of which are for Autonomous Vehicles (8.33%), three mention blockchains (12.5%), and only one mentions smart grids (4.17%). Six works analyze Intrusion Detection Systems (25%), eight mention Denial-of-Service attacks (33.33%) and three mention MitM attacks (12.5%). Six articles mention the healthcare sector (25%), while two mention the finance sector (8.33%).

The found literature has been arranged in a taxonomy presented in Figure 2. The taxonomy, all centered around the general theme of Cybersecurity, has been arranged in three hypercategories which are not independent, but rather intersecting and alternative to each other. The first hypercategory is AI method, which delves into the category of the method employed to serve any task. This hypercategory follows the standard division of AI methods, with supervised learning (Decision Trees, Random Forest, SVM, CNN, RNN, LSTM), unsupervised learning (clustering, Autoencoders, Isolation Forest), reinforcement learning (adaptive policy optimization), generative AI (GAN, Transformers for synthetic data generation, threat simulation) and hybrid approaches (AI models + fuzzy logic, blockchain, rule-based systems). The second hypercategory is the straightforward category of application domains, which correspond to the already mentioned Industry 4.0 (cyber-physical production systems, predictive maintenance, industrial IoT), healthcare infrastructure (medical IoT, electronic health records, diagnostic imaging systems), Finance (fraud detection, risk assessment, decentralized finance systems), Autonomous Vehicles (self-driving systems, vehicle-to-everything (V2X) communication, predictive maintenance), Smart Grids (false data injection detection, load forecasting, grid stability monitoring), blockchain-enabled systems (supply chain provenance, secure distributed ledgers) and wireless sensors and Cyber-Physical Systems (environmental monitoring, industrial automation, real-time control). Finally, the third and last hypercategory is that of functions, which groups works based on the type of benefit they give to the system, mainly classified based on the type of cyber attacks they tackle. They have been divided into threat detection and classification (Intrusion Detection Systems, malware classification, phishing detection), endpoint and network security (anomaly detection in IoT devices, smart grid protection, network traffic analysis), fraud and abuse prevention (transaction anomaly detection, identity verification, behavior analysis), attack mitigation and response (automated incident response, adversarial defense, zero-day mitigation) and system hardening and data protection (encryption, blockchain integration, privacy-preserving AI).

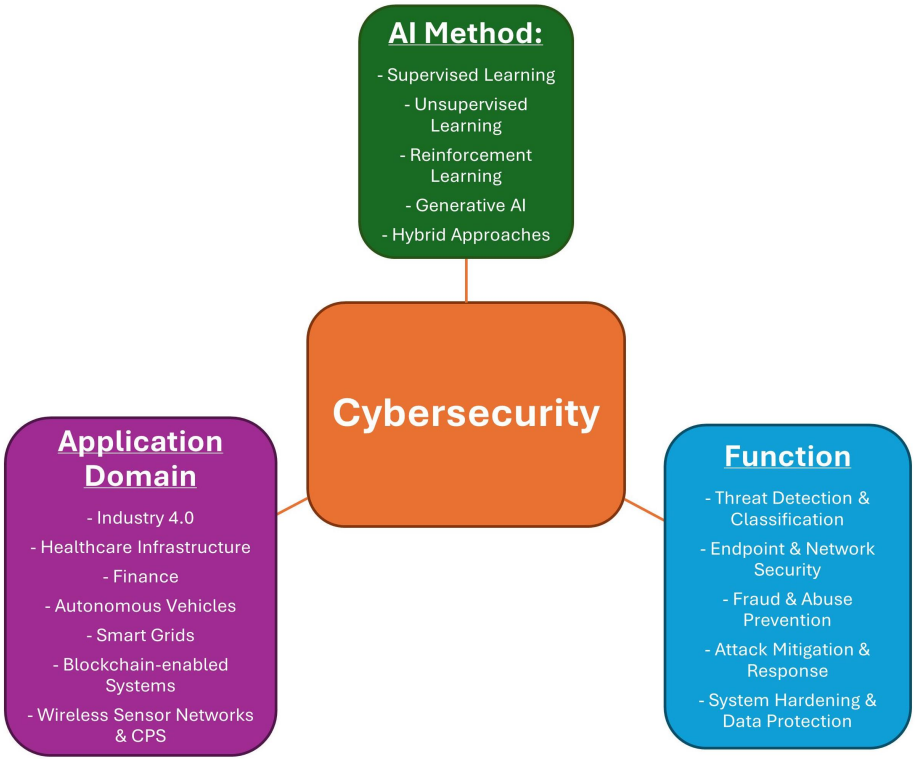

Figure 2. Taxonomy of AI techniques for cybersecurity according to the results of our PRISMA-ScR work selection. Three thematic areas have been identified: AI methods, application domains and function. The three areas are not independent from each other, but present frequent intersections, and are therefore to be intended as interchangeable classifications.

3.1. State of the Art of AI in Cybersecurity

The paper by Polito et al. [20] explores the integration of AI in cybersecurity, highlighting both its opportunities and challenges. As described by the authors, the use of AI can automate recurring tasks, identify anomalies and trends, create threat scenarios for training and simulation, and predict future threats based on historical data (past data on which to base training). However, AI can also be exploited by cybercriminals to perform more sophisticated and malicious attacks, such as through data poisoning. Furthermore, the democratization of hacking capabilities through AI makes it easier for individuals with limited technical skills to perform more sophisticated and dangerous attacks. The authors mention the use of ML and DL models, especially large language models (LLMs) such as ChatGPT and DALL-E. These models use neural networks to learn the patterns and structure of human language, generating novel and original content. The authors do not specify a particular dataset used for the research, but discuss the importance of having secure datasets for training ML algorithms. The need for data libraries and software with cybersecurity pedigrees to ensure the security of AI systems is also highlighted. Found limitations include the difficulty of fully testing and validating AI systems against all possible perturbations, the need for human control mechanisms (human supervision), such as "kill switches", to ensure the safety and reliability of AI systems, and the regulation of the use of AI, especially generative AI, which requires consensus among various stakeholders to mitigate risks and ensure transparency. Despite the challenges, the authors conclude that AI has great potential to improve cybersecurity, but it is essential to develop ad hoc cybersecurity practices and have rigorous control over the datasets and AI models used to maximize their effectiveness and efficiency, especially on attacks.

Compared to the article by Polito et al. [20], the article by Mohamed et al. [9] is instead an in-depth review of the current use of AI and ML in cybersecurity, with an analysis

of both supervised and DL algorithms. The authors cover intrusion detection, malware, network security, automation, threat intelligence, vulnerability management, and security training, presenting a complete overview of the main cybersecurity criticalities. Among the ML algorithms presented by the author are several approaches from the main classes of ML solutions, such as supervised learning (RF, SVM, neural networks), unsupervised learning (clustering, anomaly detection), DL (CNN, RNN, AE), reinforcement learning for decision optimization, and Natural Language Processing (NLP) for threat intelligence and training. From the analyses reported in the article, it is possible to observe that 45% of organizations have already implemented AI/ML in cybersecurity, mainly for intrusion detection and network security. Most of the reported cyberattacks are network intrusion attempts (20%), with other renowned frequent attacks being ransomware (62%), IoT attacks (66%), malware (43%), and Encrypted threats (4%). These quantitative data come from industrial reports and surveys without a real experimental dataset. The authors in the article highlight some limitations of ML and AI models, in particular the high computational requirement, the biases in the data and algorithms, and the problem of difficulty in interpreting the models used. Furthermore, these models have high costs, and to date, there is still a poor technological understanding.

Similarly, the paper by Wazid et al. [21] also explores the integration between cybersecurity and ML, highlighting how this synergy can improve the protection of digital systems and the effectiveness of ML models and focusing on the role of DL models (CNN and clustering). Two main approaches are analyzed: the use of ML to strengthen cybersecurity (e.g., intrusion detection) and the use of cybersecurity to protect ML models from attacks (e.g., data poisoning or privacy violations). The scope of application includes IoT environments, healthcare systems, and critical infrastructures, where protection from attacks such as Denial of Service (DoS), malware, spoofing, and attacks on ML models is crucial. The effectiveness of approaches such as CNN, clustering, blockchain, and lightweight models for malware detection is analyzed through comparative tables. As highlighted by the authors, Kumar's ML + Blockchain method achieved an accuracy of 98%, while the EveDroid system achieved an F1-score of 99%. Finally, Nguyen's CNN method achieved an accuracy of 92% and an F1-score of 94%. The datasets used to train ML models are not specified in detail, but reference is made to pre-processed data for intrusion and malware detection, often managed via cloud to exploit high computational resources. Several advantages of union between ML and cybersecurity are evidenced, such as greater accuracy in the detection of zero-day attacks, reduction in human intervention, improvement of the performance of security systems, and protection of ML models. However, the authors also highlight several limitations: compatibility issues between algorithms, overload of system resources, errors in datasets that compromise accuracy and increase vulnerability in security protocols. To tackle them, future research directions are suggested within the discussion, including the development of more robust protocols against zero-day attacks, lightweight algorithms for resource-constrained environments, and methods to improve the accuracy of ML models.

The article by Radanliev et al. [22] focuses instead on the concept of "cyber diplomacy", i.e., the use of diplomatic tools to address the challenges and exploit the opportunities offered by emerging technologies, such as AI, IoT, blockchains and quantum computing. The authors explore how these tools can facilitate international cooperation, the definition of shared norms and the resolution of conflicts in cyberspace, highlighting the importance of global agreements and standards for effective digital governance. The scope of the article is broad and intersects both the sector of national cyber security and that of international relations, emphasizing the need for multilateral coordination. The authors do not use their own experimental data, but make use of a heterogeneous set of secondary sources, such

as government policy documents, international treaties, and academic reports, organized and summarized in tables that compare different initiatives and strategies at a global level. The methodological approach is based on a systematic literature review and case studies, integrating qualitative and comparative analysis to define a theoretical framework that links emerging technologies to diplomatic dynamics. No traditional computational algorithm is present, favoring instead a structured process of qualitative analysis that adopts mapping tools, such as network diagrams, to visualize interrelationships. The results show the temporal distribution of cybersecurity commitments (with evidence from United States and United Kingdom national strategies) and the number of relevant agreements and initiatives on the international scene. The study underlines the importance of collaboration between the public and private sectors, essential to improve resilience against sophisticated attacks. Among the main critical issues, the authors highlight the difficulties in attributing attacks in a global scenario, the speed of technological changes, and the heterogeneity of regulatory systems, which place limits on the effectiveness of cyber diplomacy. Furthermore, it is reported that traditional methodologies can be slow in responding to the rapid evolution of cyberspace, requiring a continuous adaptation of diplomatic strategies. In conclusion, a joint international commitment that integrates technological innovation and modernization of diplomatic practices is required to ensure digital security, while also recognizing limitations related to the rapid evolution of threats and regulatory challenges.

3.2. Industry 4.0

The article by Yu et al. [23], instead of focusing on the synergy between cybersecurity and ML, aims to be an in-depth review on the use of ML to strengthen cyber resilience in modern industrial contexts, known as Industry 4.0. In this scenario, characterized by a strong integration between cyber-physical systems, IoT devices, and advanced automation, cybersecurity plays a crucial role. The authors analyze how ML techniques can help improve the ability of industrial organizations to prevent, detect and respond effectively to cyber attacks. The author focuses on several applications of ML, including predictive risk assessment, intrusion detection, automated incident response, threat intelligence sharing, and protection of ML models themselves from adversarial attacks. Although not using a single, specific dataset, the work includes labeled and unlabeled data, system logs, network traffic, and artificially generated synthetic data to simulate attack scenarios. These data are used to train models that can recognize anomalous or malicious behaviors in complex industrial environments. In particular, the authors employed supervised learning methods such as deep neural networks, SVMs, and decision trees, unsupervised techniques such as clustering and autoencoders, and semi-supervised approaches that combine labeled and unlabeled data. Furthermore, the use of reinforcement learning to develop adaptive security policies and dynamic responses to attacks is discussed. The results highlight how integrating ML into industrial security systems enables earlier and more accurate threat detection, increased automation in incident response, and improved adaptation to emerging threats. In particular, ML models prove effective in recognizing sophisticated attacks such as advanced persistent attacks (APTs), malware, data exfiltration, and DoS attacks, thus helping to maintain business continuity and data security. However, the authors do not overlook the limitations of these approaches. Among the main critical issues, the authors cite the vulnerability of ML models to adversarial attacks, the difficulty in interpreting the decisions of complex models, the need to have large amounts of high-quality data, and the management of privacy in data collection and sharing processes. The importance of developing robust, transparent and adaptable models, capable of operating effectively even in hostile and constantly evolving environments, is also highlighted.

Similarly, the paper by Al-Quayed et al. [24] focuses on applications of AI in cybersecurity to Industry 4.0. Instead of focusing on predictive risk and dynamic response to cyber-attacks, the authors propose a predictive framework for intrusion detection and prevention in Wireless Sensor Networks (WSNs). The system is designed to address the inherent vulnerabilities of wireless networks, which, despite being essential for real-time monitoring and management, are exposed to cyber attacks. In particular, the framework aims to identify and classify attacks such as Blackhole, Grayhole, Flooding and Scheduling, as well as distinguish normal traffic. To achieve this goal, ML and DL algorithms are integrated: a DT and a Multilayer Perceptron (MLP) for multidimensional classification and an AE for binary classification. The dataset used is WSN-DS, which replicates DoS attacks on the Low-Energy Adaptive Clustering Hierarchy (LEACH) protocol. It contains labeled data related to four types of attacks and normal behaviors. The data is cleaned, normalized, feature engineered, and then divided into training (80%) and testing (20%). The implemented models achieved very high performances; the DT obtained an accuracy of 99.48%, precision of 99.49%, recall of 99.48% and F1-score of 99.49%. The MLP showed similar metrics, with all indicators around 99.5%. The AE, used for binary classification, reported an accuracy of 91%, precision of 92%, recall of 91%, and F1-score of 91%. The benchmarks used for comparison, RF for multidimensional classification and Linear Regression (LR) for binary classification, both show lower performances than the proposed models. The framework also includes an intelligent prioritization system that orders threats based on their importance and specific impact for Industry 4.0 applications, allowing proactive and targeted interventions. The results of this system are validated through metrics such as accuracy, precision, recall, F1-score, specificity, ROC curves, and precision–recall, demonstrating its robustness. Among the limitations highlighted, the authors underline that the study is confined to the domain of WSNs in Industry 4.0 and specific to the LEACH protocol, requiring further analysis to extend the applicability to other network contexts. The need for continuous retraining of the models and the integration of dynamic threat intelligence tools is also highlighted, which entails challenges in terms of computational resources and dataset updating. The authors suggest exploring hybrid solutions and integrating international security standards (ISO/IEC 27001 [25], NIST [26], ISA/IEC 62443 [27]) to extend the effectiveness of the framework.

3.3. Finance

The work by Mishra et al. [18] focuses on analyzing the main gaps in finance by implementing AI algorithms to increase network security and ensure the identification of threats. The authors employed a Enhanced Encryption Standard (EES) to encrypt and decrypt sensitive financial information such as revenue or assets and KNN algorithm to predict if a malware attack is being performed. To test the model, the authors used a dataset on cyber attacks called Cyber Incidents 2005 to 2020, which contains around 250,000 sets of attack and defense strategies. Authors compared their EES + KNN pipeline, called Cyber Security in Financial Sector Management (CS-FSM), with other classic ML approaches such as SVM, Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). To evaluate the performances of the model, the authors selected a series of metrics typical of the financial domain, the main of which is risk reduction ratio. A comparison of risk reduction ratio showed an increase of up to 98.7% in CS-FSM compared to other approaches such as SVM (79.5%) or PCA (89.6%), showing the efficacy of the model compared to renowned ML approaches. Other metrics, such as data privacy, scalability and attack avoidance show a consistent improvement of 11-18% compared to other approaches, further confirming the efficacy of the method. The authors do however note that the scale

of the experiment was particularly small and that the analysis is narrow; therefore, further studies in this direction must consolidate the obtained results.

The work by Deshpande [28], instead, focuses on the identification of AI-related cyber threats in finance with the use of Isolation Forest (IF) for threat identification and Deep Q-Network (DQN) for adaptive response to attacks. No dataset is mentioned, but the work delineates the steps taken for data preprocessing. The performance of the model is evaluated through risk reduction compared to the unsecured system like in the work by Mishra et al. [18], with an increase of 18% (unsecured: 75%, IF + DQN: 93%). The author also mentions a decrease of 15% in false positive rate (unsecured: 20%, IF + DQN: 5%) and a $\times 3$ reduction in response time (unsecured: 30 s, IF + DQN: 10 s), making the system overall more secure with the addition of security protocols. The work, however, still presents several gaps, mainly in the absence of more information on the training data, such as dataset size or variability.

3.4. Intrusion Detection Systems

Geetha et al. [29] examine the growing threat posed by various types of cyber attacks, from hardware- to network- and application-level attacks, e.g., the role of IDS (Intrusion Detection Systems) in dealing with them [30]. The main application area is the protection of IT infrastructures, focusing on solutions to detect attacks such as DoS, phishing, malware, Structured Query Language (SQL) injection, and MitM attacks. At the methodological level, traditional ML techniques (SVM, KNN, LR, and RF) are classified and compared with DL methodologies, which are based on deep neural networks structured in multiple layers, such as Fully connected Feedforward Networks, CNNs, and RNNs. In addition, the article delves into unsupervised learning methods such as Deep Belief Networks (DBNs) and Stacked Autoencoders (SAEs), highlighting how these approaches can extract relevant features from data without a strong dependence on labeled datasets. The authors refer to classic benchmarks such as KDD Cup 99 and NSL-KDD, which typically contain network traffic records, each characterized by a set of 41 features based on Transmission Control Protocol (TCP) connection metrics (duration, protocol, service, flags, etc.). The NSL-KDD, in particular, contains about 126,000 training connections and 22,544 test connections, and includes different categories of attacks such as DoS, probing, Remote to Local (R2L), and User to Root (U2R). The richness of these structures allows testing the classification capabilities of the algorithms. The analysis of the results shows that, in several studies, some ML and DL techniques achieve very high accuracy rates, in the order of 95–99%, especially in the most common attack detection tasks. However, some limitations also emerge; real-time scalability and the management of high-speed streaming data represent significant challenges, as well as the problem of class imbalance and the difficulty in recognizing zero-day or extremely sophisticated attacks. Another criticality highlighted concerns the high computational load and the need for large hardware resources, particularly in DL models, in addition to the complexity in the feature extraction phase, despite the use of unsupervised learning techniques. Furthermore, the ecosystem of dedicated platforms (such as TensorFlow, Keras and Apache Spark MLlib) that facilitate the implementation of such algorithms is discussed, highlighting how the need for labeled datasets and the complexity of the models can limit the effectiveness of some approaches in real-world scenarios.

Similarly, the work by Halbouni et al. [31] explores ML and DL approaches applied to cybersecurity with a focus on IDS. In particular, the authors focus on the detection of cyber attacks in network environments, IoT, and critical infrastructures, highlighting the value of IDS in protecting data and systems against unauthorized access. Among the application areas covered are the security of corporate networks and cloud environments where the rapid evolution of cyber-attacks requires predictive and adaptive methods. The

authors illustrate the use of several known datasets in IDS, among which are the already mentioned KDD Cup 1999 and its evolution, NSL-KDD. More recent datasets such as UNSW-IDS15, collected in a laboratory environment and divided into nine families of attacks, and CIC-IDS2017, with 84 features, which represents both real and emulated traffic, are also analyzed. The structure of these datasets includes traffic records enriched with numerical and categorical attributes, useful for describing the behavior of connections and classifying the types of attacks. The article compares several ML algorithms, such as SVM, RF, DT, and KNN, with DL techniques based on deep neural networks such as CNN, RNN, LSTM, AE, and DBN. The ability of DL models to automatically perform feature extraction is highlighted, reducing manual intervention compared to traditional ML approaches. The reported results demonstrate that, with an adequate amount of data, DL models often achieve high accuracies (in some cases over 99%) and low false alarm rates. However, the authors note that the performance of the models can vary significantly depending on the dataset and the types of attacks considered. Among the limitations highlighted is the use of obsolete or redundant datasets, which may not reflect the dynamics of real traffic. Such concerns align with recent findings in the field of digital image forensics, where CNN-based copy-move forgery detectors have been recently implemented [32]. The imbalance present in the datasets is also highlighted, which penalizes the detection of less represented attacks such as U2R and R2L. Another criticality consists in the high computational requirement of DL models, which require advanced hardware resources (GPU) and long training times. Finally, the low interpretability of the "black-box" of DL models is mentioned as a problem, especially in contexts in which it is essential to explain the choices made by the system.

The paper by Siva Shankar et al. [33] proposes a new IDS approach based on DL and AI techniques, integrated with optimization methods, to detect attacks in a network of systems. Compared to previous papers, it proposes a hybrid approach that integrates Multi-layer Perceptron with fuzzy logic, focusing on the protection of network infrastructures. The approach is designed for environments where resources are limited and the variety of attacks is constantly evolving. Two standard datasets are used. The first is the already mentioned NSL-KDD, divided into training data (KDDTrain+) and test data (KDDTest+), while the second is UNSW-NB15, created with the IXIA PerfectStorm application. UNSW-NB15 contains 49 attributes and records nine types of attacks (backdoor, fuzzers, DoS, shellcode, reconnaissance, etc.), with well-defined training and testing sets. The proposed approach is divided into two main phases. In the first phase, the feature extraction and selection method is based on the CHO (Corporate Hierarchy Optimization) algorithm, inspired by corporate organizational structure, to identify the most informative features. Subsequently, the classification is performed using the GEO-SMPIF model, a hybrid system that integrates a self-constructing MLP with a fuzzy system, optimized through the Golden Eagle Optimization algorithm for hyperparameter tuning. Data is first pre-processed (transformation, normalization via Min-Max) to convert non-numeric attributes into numerical values suitable for training, and then used to train the GEO-SMPIF model. The GEO-SMPIF structure includes a fuzzification layer to convert inputs into fuzzy sets, intermediate layers representing the fuzzy inference system, hidden layers for the DL section, and finally a defuzzification layer to obtain the final output (classification into "normal" or "attack"). The results on the NSL-KDD and UNSW-NB15 datasets indicate an accuracy of 99.99% (NSL-KDD) and 99.97% (UNSW-NB15), with a detection rate close to 100% and extremely low false alarms (FPR of 0.04 and 0.065, respectively). The authors compare the new approach with existing methods (DAE+DNN, SVM-WOA, EFSGOA), highlighting improvements in terms of accuracy, sensitivity, and computation time. However, some potential critical issues arise, including the risk of overfitting, particularly in the

NSL-KDD dataset. The challenge of maintaining good generalization in the presence of novel attacks or unbalanced data is also highlighted.

The work by Nabi et al. [34], compared to other works which relied on DL methods, studies the improvement of IDS through dimensionality reduction techniques, with a comparative focus on two approaches: PCA and Random Projection (RP). The goal is to reduce training time and improve the accuracy of classifiers, while minimizing false alarms. Once again, the study uses the NSL-KDD dataset to test the proposed approaches. The researchers employ five supervised learning algorithms in comparison: BayesNet, Naïve Bayes, J48, Partial Decision Tree (PART), and RF. Initially, they were trained on the complete set of features, obtaining, for example, with J48 an accuracy of 79.1% and a false positive rate of 18.5%. Subsequently, PCA is applied to reduce the dimensionality and an improvement is highlighted for some algorithms (Naïve Bayes reaches 78.8% accuracy), However, the main contribution of the paper emerges with the application of RP, which, using a Gaussian random matrix, allows us to reduce the dimensions more quickly and effectively; in this context, the PART algorithm achieves the best accuracy overall, equal to 82.0%, with a better false positive rate than other methodologies. The results demonstrate that RP not only preserves the essential features of the data, but also significantly reduces the training time compared to PCA. Among the limitations, the authors highlight the difficulty in correctly classifying the "anomalous" class, which shows a lower true positive rate than the "normal" class, making the detection of new intrusions particularly challenging. Furthermore, possible overfitting issues are reported, especially for algorithms such as RF, and it is noted that, although the NSL-KDD dataset is a consolidated benchmark, its limited structure may not represent all the variations encountered in real environments. The author suggests that, for further developments, alternative dimensionality reduction techniques such as Latent Dirichlet Allocation (LDA), Kernel PCA or even DL-based methods could be explored, in order to obtain increasingly robust and efficient intrusion detection systems in evolving cyber threat scenarios.

3.5. Internet of Things Environments

The paper by Abdullahi et al. [12] is a systematic literature review that analyzes the use of AI techniques, both ML and DL, for cyber attack detection in the IoT environment. The scope of application concerns the security of IoT systems in sectors such as smart cities, healthcare, industry, and critical infrastructures, where data protection and attack prevention are key. The work highlights issues related to IoT network traffic datasets, in particular class imbalance (e.g., binary and multi-class problems) that can affect false alarm rates. Common datasets reported in research include NSL-KDD, KDD99, Bot-IoT, AWID, and other collections, which typically contain numerous attributes to describe normal and abnormal behavior. The algorithms used include supervised techniques such as SVM, RF, DT, KNN, and Naïve Bayes, as well as DL methods such as CNN, RNN, LSTM, AE, and DBN. Many studies report high accuracies, sometimes above 98–99%, especially with SVM and RF, which are particularly effective due to their ability to balance accuracy and memory management. The authors evaluate the performance of the models using metrics such as accuracy, precision, recall, and F1-score, also highlighting the trade-offs between training time and prediction speed. The paper also maps the various solutions currently proposed, classifying them based on the ML and DL algorithms employed, the IDS-based architectures, and the types of attacks detected (DoS, DDoS, Probe, R2L, U2R, etc.). A recurring criticism concerns the use of obsolete datasets (e.g., NSL-KDD and KDD99), which may not reflect the current attack dynamics in the IoT context, therefore showing a need to adopt up-to-date datasets, preferably derived from real IoT environments. Other highlighted limitations are the high computational complexity and long training times,

especially in DL models, and the difficulty in handling highly unbalanced datasets, which lead to high false positives. The study suggests that the integration of hybrid models, capable of combining multiple AI techniques, could further improve attack detection. Finally, the authors invite others to experiment with new feature selection techniques, to combine hybrid approaches and to also consider emerging methods such as transformers to address security challenges in the IoT environment.

While Abdullahi et al. [12] focus on traditional algorithms, the paper by Bhandari et al. [35] presents a middleware framework based on distributed deep neural networks for cyber-attack detection in smart IoT environments, aimed at protecting devices and networks in highly heterogeneous scenarios. The scope of application concerns cybersecurity in IoT ecosystems, where the multiplicity of devices with limited resources and heterogeneous protocols makes the implementation of protection solutions particularly complex. The proposed framework adopts a multi-agent approach, where learning operations are performed in the cloud environment while inferences are executed on edge gateways, thus reducing the computational load on IoT nodes. The datasets used for training and testing are IoT-23 and Edge-IIoTset; the first contains semi-structured captures in .pcap format with over 325 million instances (of which about 294 million are malicious samples), while the second collects about 1.36 million benign samples and 545 thousand attacks. Data is subjected to feature extraction and selection phases and transformed into a structured format suitable for training ML models. Among the techniques used, the main algorithm is a deep neural network (DNN), which is compared with other models such as SVM, RF, DT, Gradient Boosting, and Naive Bayes. The experimental results show that the DNN model achieves about 93% accuracy and an F1-score of around 92% in the tests performed on the two datasets. On the hardware side, the framework has a limited impact; model execution results in an average bandwidth increase of less than 30 kb/s and a 2% increase in CPU usage. Memory requirements are moderate, with about 0.42 GB used on NVIDIA Jetson devices and 0.2 GB on Raspberry Pi, while power consumption shows an average increase of 13.5% per device with the model active. The system has been tested in several real-world use cases, such as smart waste bins, temperature sensors, passenger counters, and weather stations, demonstrating its adaptability to diverse scenarios. The authors note that, although the DNN offers the best overall performance, some results are lower on the IoT-23 dataset due to significant imbalances in attack classes. This issue, highlighted for example by the presence of only two or three instances for some attacks (as in the case of C-Mirai), can lead to overfitting phenomena. Among the limitations, the artificial nature of the datasets employed is recognized, which may not fully reflect the dynamics of real attacks and presents marked inequalities in the number of samples for each attack category. The authors suggest, as a future research direction, the adoption of few-shot learning techniques to counteract the problem of data imbalances, in addition to the implementation of lighter models in C language to save resources. Finally, the study paves the way for the integration of additional data sources, such as logs, reports, and information from sensors, to further improve the system's ability to identify anomalies and detect attacks. Furthermore, recent research has demonstrated that compressing recurrent architectures such as LSTM via Singular Value Decomposition (SVD) enables real-time inference on resource-constrained platforms [36].

Similarly, the work by Becerra-Suarez et al. [37] evaluates the performance of different DL models for the classification of cybersecurity attacks in IoT networks using an updated and realistic dataset. In particular, the study focuses on the analysis of the capabilities of models based on DNN, LSTM, and CNN to identify attacks such as Distributed Denial-of-Service (DDoS), DoS, Mirai, recon, spoofing, brute force, and web attacks in IoT environments. Once again, the focus is on the security of IoT ecosystems, where the

growing number of connected devices and intrinsic vulnerabilities makes timely detection of cyberattacks essential. The dataset used, called CICIoT2023 and developed by the Canadian Institute for Cybersecurity, includes over 46 million records and 47 features, which represent events derived from windows of communication packets between two hosts in an IoT topology with 105 devices. To reduce the computational load, sampling was performed using only 1% of the records for each category, and after eliminating irrelevant features (such as those consisting only of zeros), the final dataset was simplified to 40 features. The DL models underwent a thorough preprocessing, feature selection and data standardization process. The DNN model architecture is based on multiple dense layers with ReLU activations, while the LSTM model uses two LSTM layers to capture potential temporal dependencies, although in this context, the LSTM did not offer good results due to the local nature of the data. The CNN model, instead, uses multiple convolutions with kernels of different sizes (3, 5, and 11) to capture patterns at different scales, followed by an additional convolution and max-pooling operations, proving to be particularly effective. The evaluation metrics (accuracy, precision, recall, and F1 score) show that the CNN model achieved an accuracy of 99.10% in multiclass classification and 99.40% in binary classification, with significantly reduced training and inference times compared to the other models. The results also show that, although the LSTM model is typically suitable for sequential data, in this case, its performance was lower, since the 40 features of the dataset do not present significant temporal dependencies. On the contrary, the DNN achieved very good performances but still lower than those of the CNN, which instead manages to more effectively extract features relevant for distinguishing attacks. Among the limitations mentioned, there is the need for further investigations to evaluate other architectures and to test the model on different datasets, in addition to the computational complexity that, although reduced in CNN, could increase with the addition of further features or in more complex scenarios.

The work by Bendiab et al. [38] presents a systematic analysis of AI + Blockchain for the protection of autonomous vehicle (AV) systems. Similarly to other works, such as the ones by Polito et al. [20] or Mohamed et al. [9], Bendiab et al. do not present a novel pipeline, but rather analyze the current state of the art in the application of AI and Blockchain in the AV sector and define future research lines in this field. What emerged from their research is that AVs can be likened to IoT systems and Cyber-Physical Systems (CPS) due to the interconnection between the parts which compose the full system, such as sensors and control units, and the connection to external units such as other vehicles. For this reason, cybersecurity is particularly important in such a scenario, and Blockchain can guarantee secure data storage and protected communication with a Proof-of-Work (PoW) and Proof-of-Stake (PoS) consensus mechanism, ensuring possible uses even in the forensic field to prevent malicious tampering when determining accident responsibility. At the same time, the introduction of AI in the form of KNN, SVM, Hidden Markov Models (HMMs) or DL can favor Intrusion Detection by identifying attacks such as DDoS, spoofing or malware, allowing even malware classification with up to 99.9% accuracy and further ensuring data security with federated learning. The authors note several limitations in current literature, such as scalability, computational performance, lack of datasets and protocols, and most importantly a lack of transparency, which can be tackled with Explainable AI (xAI), suggesting that future research should focus on the integration of quantum computing and federated learning in current pipelines.

The work of Aldhyani et al. [39], instead, presents a defined solution for intrusion detection in AVs through a hybrid CNN + LSTM network. The authors constructed a dataset from real Controller Area Network (CAN) traffic data which includes spoofing, flood and replaying attacks manually injected every 3 to 5 s. This dataset was used to train a simple network composed of a temporal LSTM block for long-range dependencies and a

three-block CNN for spatial dependencies. The full network aims at classifying IDS, with the performances evaluated with accuracy, F1 score, precision and recall. While a simple CNN achieved 86% overall accuracy, but failing to detect any attack packet, the complete CNN+LSTM model shows a relevant capability to identify attacks, with an accuracy of 95.44% over the full labels (benign, flood, fuzzing, replaying, spoofing) and of 97.3% over a reduced dataset (benign, flood, fuzzing). However, the poor results obtained in the removed classes (replaying and spoofing) show that this direction would still benefit for further experimentation, especially due to the statistical imbalance found by the authors between labels and features in the employed dataset.

3.6. Attack Detection and Data Protection

Gaba et al. [40] systematically analyzed the application of DL techniques to improve the security of CPS, focusing on the detection of cyber attacks. The scope of application covers critical sectors such as energy, industry, healthcare, and essential infrastructures, where the integration of IoT devices increases the attack surface. A mathematical framework is proposed that reduces the overall cost associated with vulnerabilities, modeling the system as a set of components and vulnerabilities. The framework uses a threat vector to identify potential attacks and to guide the mitigation of criticalities. The adopted dataset is structured as a collection of (Input, Label) pairs and combines data from simulations and real testbeds, with some well-known datasets, such as the CICIDS2017 and the SMACK dataset, employed to address the complexity and heterogeneity of CPS data. The architectures used include CNN, RNN, AE and GANs, capable of extracting features from data in a multi-layered way. These models allow us to identify anomalies and atypical behaviors, signals of potential attacks in CPS. The results show a clear improvement in attack detection compared to traditional methods. However, the work highlights limitations such as the scarcity of labeled data and the difficulties of scalability in real-time environments. Further critical issues concern the lack of interpretability of the models and their vulnerability to adversarial attacks. To strengthen security, the study suggests the adoption of techniques such as transfer learning and emerging paradigms (e.g., self-supervised learning), as well as the development of more diverse datasets.

The paper by Obonna et al. [41] addresses the problem of detecting MitM attacks in process control networks (PCNs) in the Oil and Gas sector, highlighting how the growing integration between Information Technology (IT) and Operational Technology (OT) technologies has exposed these systems to ever greater cyber risks. The application area is that of critical infrastructures, where the security of SCADA systems is essential to avoid outages, catastrophic accidents, and damage to the environment. For the study, a real dataset composed of 68,722 pressure measurements taken from a three-phase separation plant, collected by a SCADA system, was used; these data were divided into training (60%), testing (25%), and validation (15%). Furthermore, to validate the results, three public datasets were used: WUSTL-2018, ORNL Power Grid, and TON IoT. On the main dataset, the authors applied different ML algorithms, such as IF, KNN, Local Outlier Factor (LOF), LSTM, and various decision tree-based models (e.g., fine, medium, coarse tree, and ensemble methods such as bagged tree). The primary goal was to detect anomalies in the data due to MitM attacks, measuring metrics such as accuracy, false positive rate (FPR), training time, and prediction speed. In particular, the coarse tree model achieved outstanding performances, with 100% accuracy, training time of about 0.4549 ms and extremely fast prediction, demonstrating an effective handling of attacks with close to zero false positives. Numerical results also highlight that, using ensemble techniques such as bagged trees on public datasets, high performances can be achieved, although at the cost of longer training times. Some models, such as those based on LSTM or KNN, have instead

shown criticalities in detecting extreme anomalies or a high false alarm rate, highlighting the trade-off between accuracy and computational complexity. Among the limitations, the authors report the difficulty in distinguishing similar attacks (e.g., MitM vs. backdoor) and the need for further tuning to optimize detection in multiclass scenarios, in addition to the fact that the specific dataset may not fully represent the dynamics of real environments.

The paper by Shees et al. [42] addresses the issue of cybersecurity in smart grids, focusing in particular on the so-called "false data injection attacks" (FDIAs). In an era in which the integration of advanced technologies and data-driven systems in electric grids has brought enormous benefits in terms of efficiency and sustainability, vulnerability to such attacks becomes crucial because data manipulation in control systems can compromise the entire energy distribution infrastructure. To analyze this problem, the authors used a dataset from the Oak Ridge National Laboratory, which collects numerous event scenarios, divided into natural events, normal events, and attacks. Interestingly, the dataset is composed of 128 features, 29 variables for each of the four Phasor Measurement Units (PMUs), and an additional 12 features derived from system logs, thus providing a detailed mapping of the operational situation of the electric system. The investigation strategy is based on the application and comparison of six ML models, namely Extra Tree Classifier, RF, Extreme Gradient Boosting (XGBoost), LR, DT, and Bagging Classifier. These algorithms have been trained and tested using optimization techniques such as Bayesian optimization and validated through a 70/30 split of the dataset accompanied by a 5-fold cross-validation to limit the risk of overfitting. The results showed a clear difference in performance. The approaches based on ensemble models, such as Extra Tree, RF, and XGBoost, offered high accuracies (98%, 97%, and 97%, respectively) and good trade-offs between precision and recall, highlighting a greater ability to capture the complex non-linear relationships present in the data. On the contrary, LR, due to its intrinsically linear nature, showed significantly lower results, demonstrating difficulties in correctly distinguishing between normal operations and attacks. A critical aspect that emerged concerns the difficulty in distinguishing low-impact attacks from normal operating processes, as the related variations in measurements (such as voltage and current) are minimal and easily confounded. This issue leads to a higher risk of classification errors, especially for less sophisticated models. The authors also highlight that, although ensemble algorithms show training scores of 1.00, there is a risk of overfitting, which is mitigated through adequate cross-validation procedures but could still be present. The authors suggest that the adoption of more advanced techniques, such as DL architectures or federated learning methods, could further strengthen the resilience of smart grids against this kind of threat, paving the way for real-time implementations in large-scale scenarios.

Shaukat et al.'s [43] paper focuses on the application of ML techniques to tackle cyber threats, especially intrusion, spam, and malware detection. In this context, various algorithms such as SVM, DT, RF, Artificial Neural Networks (ANNs), DBN, and Naïve Bayes are compared, analyzing both their performance and computational complexity. The datasets used include KDD Cup 99, NSL-KDD, and DARPA for intrusion detection, along with collections such as Enron, Spambase, SMS collections, and Twitter datasets for spam detection. For malware, static, dynamic, and even custom datasets such as VirusShare are used. Numerical results show significant performance, with SVM reaching an accuracy of up to 99.30% on KDD Cup 99, DT recording performances around 99.96% in some configurations and DBN generally above 97.50% for IDS. RF and Naïve Bayes also show high precision and recall values, depending on the dataset and the sub-domain considered. The authors, however, point out several limitations, such as the scarcity of updated and balanced datasets, the variation of results related to the choice of feature extraction methods, and the high computational cost of some models. Furthermore, it is

highlighted how the lack of standardization in metrics (some studies report only accuracy while others include precision and recall) complicates the comparison. The vulnerability of algorithms to adversarial inputs and the need for robust models in applications where speed and reliability are critical are also discussed.

The article by Teo et al. [44] examines the topic of cybersecurity in the era of Generative Artificial Intelligence (GenAI), focusing on the implications for data protection and security in the healthcare sector. Compared to previous works, which mainly analyzed older DL models, Teo et al. [44] focus instead on newer models such as Generative Adversarial Networks (GAN) and Transformers. An overview of GenAI technologies is presented, highlighting their potential to generate original content through learning from data. The scope of application is, as mentioned, mainly the healthcare sector, with specific references to clinical, operational, and research cases, including a particular focus on ophthalmology. As explained in the article, the training of models occurs through the analysis of large unstructured datasets, composed of textual data, diagnostic images, and other forms of clinical information. While no specific dataset is described, the importance of using both real and synthetic data to train and evaluate GenAI models is highlighted. The authors illustrate the dynamics of GANs, where a generator network competes with a discriminator to produce realistic outputs, and Transformers, which instead analyze complex relationships in sequential data through their attention mechanisms. LLMs are also cited in the work. The article demonstrates how these algorithms have revolutionized tasks such as image generation, text synthesis and clinical document processing, but also discusses how, despite the potential in terms of efficiency and innovation, GenAI models are subject to concretely relevant risks such as data poisoning, model inversion and membership inference attacks. A further problem concerns the propagation of bias and the generation of hallucinations, i.e., plausible but completely invented information. Among possible mitigation strategies, the authors discuss the use of cryptographic techniques, differential privacy, and federated learning to protect sensitive data. Once again, the poor explainability of the models, which operate as "black boxes", is cited. In case of clinical errors, it is particularly relevant as it complicates tracing decisions and managing responsibilities. Finally, the need for interdisciplinary collaboration and the development of regulatory standards that allow the safe exploitation of the potential of GenAI in the healthcare sector is highlighted.

3.7. Explainable AI: Beyond the Black Box

The article by Pawlicki et al. [45] thoroughly and systematically address future possibilities and research directions in the field of xAI applied to cybersecurity, with a particular focus on intrusion detection systems that use DL and AI techniques as also discussed in the article by Geetha et al. [29] The authors highlight how the need for transparency in decision-making models, often considered "black-box", is crucial in high-risk sectors, where wrong decisions could have serious consequences, such as the previously mentioned healthcare sector. For this reason, the discussion focuses not only on xAI techniques that can make model choices understandable, but also on the need for user-centered approaches, able to provide contextually relevant explanations that are easily interpretable by non-specialists. The authors discuss numerous XAI algorithms and methodologies, both at local and global levels. Among these, methods such as Local Interpretable Model-agnostic Explanations (LIME) and anchors offer precise explanations for specific decisions, while techniques such as decision trees, rule lists, Partial Dependence Plots, and Shapley Additive exPlanations (SHAP) provide a more comprehensive view of the decision-making process of models. The visualizations and numerical measurements reported highlight, for example, how the analyzed models show particular values of precision. A central aspect also concerns the dataset used for the analysis; it is not data from a single source,

but rather a heterogeneous set of publications that highlight current and future trends in research, with a temporal distribution that underlines the growth of interest in the topic in recent years. Numerical results, such as the number of articles per database and per year (with a significant concentration in 2023), reinforce the idea of a rapidly evolving field. The work does not fail to highlight current challenges and limitations; there is a widespread lack of standardized metrics to evaluate the interpretability of systems, and a balance between model complexity, accuracy, and transparency is still difficult to achieve. At the same time, the need to integrate interdisciplinary approaches involving humanities and social sciences to better understand the interaction between humans and machines is highlighted, as well as the importance of developing interactive and hybrid methods capable of adapting to the needs of a wide range of users. Finally, the authors orient the future of research in xAI for cybersecurity towards the development of interfaces that offer clear and effective explanations, increasing user trust and promoting more informed decisions in environments characterized by high criticality and a continuous evolution of cyber threats.

While Pawlicki et al. [45] focus on intrusion detection, the article by Lee et al. [46] addresses the problem of low interpretability of AI systems and xAI used for anomaly detection in cybersecurity, proposing a framework that supports analysts' decisions through clear and targeted explanations. In this context, in fact, the black-box nature of traditional models makes it difficult to understand the reasons behind a given prediction, especially when it comes to discriminating between normal situations and attacks. The framework is based on the generation of "References", i.e., reference data that, although belonging to the normal category, are chosen for their similarity with the anomalous data to be interpreted. This methodology allows for a direct comparison (Feature Value Comparison) between the target data and the reference data, thus offering intuitive explanations on the detected anomalies. The system is tested on logs collected in Endpoint Detection and Response (EDR) environments. These logs contain essential information (such as unique process ID, parent process data, name, and execution) and are preprocessed using 2-g and hashing tokenization techniques, generating numerical embeddings of all features. The framework, composed of an AE model for anomaly detection, showed excellent performances. On an HWP process classified as an attack, the system recorded a Root Mean Square Error (RMSE) of about 0.51118 and a p-Value of 0.0704, while a false alarm case showed significantly lower values (RMSE and p-Value significantly higher and lower than expected, respectively). Such quantitative differences facilitate the identification of real attacks versus false positives. The authors point out, however, that the use of the mean in the Mean Square Error (MSE) could "smooth out" anomalies concentrated in a few features, suggesting the adoption of specific weights to emphasize such discrepancies. Furthermore, while demonstrating numerous advantages in reducing false alarms and increasing interpretation effectiveness, the framework needs further refinements to adapt to different unsupervised learning models and to address the variability of anomalous patterns.

3.8. Summary of Results

Table 2 summarizes the findings of the discussed works, providing a schematized representation of each work, the datasets employed, the algorithms used, and the best results obtained by each study. Works which are only theoretical, and therefore do not include numerical results, have not been included in the table. For this reason, of the 24 selected works, only 15 have been included in the table. As evidenced, the majority of works use accuracy as evaluation metric (the only exception is the work by Lee et al. [46] and the two works on finance by Mishra et al. [18] and Deshpande [28]), allowing for a mostly uniform comparison of the obtained results. The considered works show a

consistent accuracy > 90%, with the only outlier being the work by Nabi et al. [34], which has a much lower 82% accuracy. The majority have shown an accuracy > 97%, with some even exceeding 99%, such as the works by Becerra-Suarez et al. [37] and the one by Shaukat et al. [43].

It must be noted that, while in terms of metrics, algorithms can easily be compared, the different domains of application, and especially the different datasets employed for each task, could make results not properly comparable. This stems mostly from an imbalance in dataset size and structure, which consequently renders some tasks more easily solvable compared to most complex ones. Similarly, datasets with a larger number of features, such as the KDD Cup 99 and NSL-KDD with their 41 features, enrich the complexity of training, leading models trained on such datasets to possibly be subject to lower accuracy while also ensuring a better understanding of the structure of data they have to analyze. However, given the similarity in results between the majority of analyzed models, we can assume this problem to be ignorable in this discussion.

At the same time, a precise comparison of the results is unfeasible due to the incompatibility of methodologies, with works focusing on different cybersecurity techniques such as intrusion detection or cyber attack classification to obtain the same result of strengthening system security. This difference makes results hardly comparable, as while the metrics may be coincident (as previously mentioned, accuracy is often the reported metric), they would derive from vastly different tasks. This heterogeneity is expected due to the various types of cyber attacks found in the literature, some of which require specific protocols. For example, attacks directed to AI models through techniques such as data poisoning or model inversion [44] require data protection algorithms, whereas other types of attacks such as DoS or phishing would benefit from intrusion detection systems [29] or identification [37].

As some authors considered in their discussion, a considerable problem could instead be overfitting, closely tied to an apparent high performance of a model, which is, however, overspecialized in recognizing samples from the training set. For this reason, it could be argued that many of the near-perfect results obtained by authors could be caused by overfitting, which is, in some cases, mitigated by the authors themselves through a careful preprocessing of training data.

As shown in Table 3, the most common category of attacks found in the literature is intrusion-based attacks. Being the most broad category of attacks, as it includes common cyber threats such as malware, backdoor and DoS, the majority of analyzed works focus on intrusion detection as their main scope. Several other attacks, such as phishing, adversarial attacks or data poisoning, have been studied, but the use of AI in cybersecurity is mostly directed towards the identification of threats and the classification of attacks, with only a small fraction of works focusing on topics such as protection and explainability.

To answer our research question, 'Is AI an effective method to enhance current infrastructures' cybersecurity?', as our literature analysis proved, AI can indeed help in the recognition and classification of cyber attacks in the context of Industry 4.0, healthcare, finance, robotics and IoT systems [47], tackling common attacks such as DoS, MitM, Mirai with accuracies up to 100% over a variety of different ML and DL methods, both lightweight and more complex. Moreover, the implementation of xAI solutions can aid sector experts by providing an analysis of the black box of AI algorithms in order to make it simpler for humans to interact with complex algorithms and verify the efficacy of their provided results.

Table 2. Table of the summary of studies included in the review which provided numerical results. The table includes authors, publication year, dataset used, employed algorithms, and best results obtained for each study.

Authors	Year	Dataset	Algorithms	Best Results
Al-Quayed et al. [24]	2024	WSN-DS (Kaggle data, attacks on LEACH protocol)	DT, MLP, AE	DT/MLP: Accuracy ∼99.5%; Autoencoder: Accuracy 91%, Precision 92%, Recall 91%, F1-score 91%
Mishra et al. [18]	2023	Cyber Incidents 2005 to 2020	EES+KNN (CS-FSM), SVM, PCA, LDA, CPS-IoT, DM, CATRAM	CS-FSM: Data privacy 96.1%, Scalability 97.2%, Risk reduction 98.7%, Data protection 95.4%, Attack avoidance 94.3%
Deshpande [28]	2024	Not mentioned	IF + DQN	Detection Rate: 93%, False Positive Rate: 5%
Geetha et al. [29]	2021	KDD Cup 99, NSL-KDD	SVM, KNN, LR, RF; FNN, CNN, RNN, DBN, SAE	Accuracy 95–99%
Halbouni et al. [31]	2022	UNSW-IDS15, CIC-IDS2017	RF, DT, KNN; CNN, LSTM, AE, DBN	Accuracy > 99% (low rate of false alarms)
Siva Shankar et al. [33]	2024	NSL-KDD, UNSW-NB15	MLP with fuzzy systems (GEO-SMPIF with Golden Eagle Optimization)	NSL-KDD: Accuracy 99.99%, FPR 0.04; UNSW-NB15: Accuracy 99.97%, FPR 0.065
Nabi et al. [34]	2024	NSL-KDD	PCA, RP; BayesNet, Naïve Bayes, J48, PART, RF	J48: Accuracy 79.1%, FPR 18.5%; PART with RP: Accuracy 82.0%
Abdullahi et al. [12]	2022	NSL-KDD, KDD Cup 99, Bot-IoT, AWID	SVM, RF, DT, KNN, Naïve Bayes; CNN, LSTM, AE, DBN	Accuracy > 98–99%
Bhandari et al. [35]	2023	IoT-23, Edge-IIoTset	DNN, SVM, RF, DT, Gradient Boosting, Naïve Bayes	DNN: Accuracy ∼93%, F1-score ∼92%; resources: band < 30 kb/s, +2% CPU, memory 0.42 GB (Jetson) vs. 0.2 GB (Raspberry Pi), Energy +13.5%
Becerra-Suarez et al. [37]	2024	CICIoT2023	DNN, LSTM, CNN	CNN: Multi-class Accuracy 99.10%, Binary Accuracy 99.40%
Aldhyani et al. [39]	2022	Real dataset on CAN traffic data	CNN + LSTM	3-class Accuracy 97.3%
Obonna et al. [41]	2023	Real dataset from SCADA plant, WUSTL-2018, ORNL Power Grid, TON IoT	IF, KNN, Local Outlier Factor, LSTM, DT ('coarse tree')	Coarse Tree: Accuracy 100%, Training Time ∼0.4549 ms, false alarms negligible
Shees et al. [42]	2024	Dataset Oak Ridge National Laboratory	Ensemble: Extra Tree Classifier, RF, XGBoost; LR, DT, Bagging Classifier	Extra Tree: Accuracy 98%; Random Forest and XGBoost: Accuracy 97%

Table 2. *Cont.*

Authors	Year	Dataset	Algorithms	Best Results
Shaukat et al. [43]	2020	KDD Cup 99, NSL-KDD, DARPA, Enron, Spambase, SMS/Twitter datasets, VirusShare, etc.	SVM, DT, RF, ANN, DBN, Naïve Bayes	SVM: Accuracy 99.30%; Decision Tree: Accuracy up to 99.96%; DBN: Accuracy > 97.50%
Lee et al. [46]	2023	Endpoint Detection and Response (EDR) system logs	XAI framework based on AE and reference-based decision support	RMSE ∼0.51118, *p*-value ∼0.0704 for attack detection

Table 3. Table of the summarization of the cyber attacks discussed in each work included in the review and their main scope of application.

Authors	Cyber Attack	Main Scope
Polito et al. [20]	Data poisoning	Identification and prediction
Mohamed et al. [9]	Ransomware, IoT attacks, malware, encrypted threats	Intrusion detection
Wazid et al. [21]	Data poisoning, privacy violation, DoS, malware, spoofing, attacks on ML, zero-day	Intrusion detection and ML protection
Radanliev et al. [22]	Not specified	Cyber diplomacy
Yu et al. [23]	Adversarial attacks, advanced persistent attacks, malware, data exfiltration, DoS	Prevention, detection and response
Al-Quayed et al. [24]	Blackhole, grayhole, flooding, scheduling	Attack classification
Mishra et al. [18]	Malware	Protection and identification
Deshpande [28]	Not specified	Identification and response
Geetha et al. [29]	DoS, phishing, malware, SQL injection, MitM	Intrusion detection
Halbouni et al. [31]	Intrusion (not specified)	Intrusion detection
Siva Shankar et al. [33]	Probing, R2L, U2R, analysis, backdoor, fuzzers, DoS, exploits generic, shellcode, reconnaissance, worms	Intrusion detection
Nabi et al. [34]	Intrusion (not specified)	Intrusion detection and classification
Abdullahi et al. [12]	DoS, DDoS, Probe, R2L, U2R	Intrusion detection and classification
Bhandari et al. [35]	Not specified	Identification
Becerra-Suarez et al. [37]	DDoS, DoS, Mirai, recon, spoofing, brute force, web attacks	Identification and classification
Bendiab et al. [38]	DDoS, spoofing, malware	Intrusion detection and classification
Aldhyani et al. [39]	Flood, fuzzing, replaying, spoofing	Intrusion detection
Gaba et al. [40]	Not specified	Identification
Obonna et al. [41]	MitM	Identification

Table 3. *Cont.*

Authors	Cyber Attack	Main Scope
Shees et al. [42]	False data injection attacks	Identification
Shaukat et al. [43]	Intrusion, spam, malware	Intrusion, spam and malware detection
Teo et al. [44]	Data poisoning, model inversion and membership inference	Data protection
Pawlicki et al. [45]	Not specified	Explainability
Lee et al. [46]	Not specified	Explainability

4. Discussion

The use of methodologies based on AI and ML offers concrete advantages in the detection and management of cyber attacks. Numerous studies have reported very high performances (some approaches exceed 99% accuracy on benchmarks such as NSL-KDD, UNSW-IDS15, and CICIDS2017) and demonstrated the ability to identify anomalies and attacks with precision, contributing to a rapid response to security incidents. The use of advanced algorithms such as deep neural networks, hybrid models integrated with fuzzy logic, and dimensionality reduction techniques has favored the improvement of performance in terms of accuracy and reduction in false alarms. ML and DL are equally used solutions, with Decision Trees and Random Forest being the most frequently used approach in the literature, while the best results are obtained by DL solutions such as Convolutional Neural Networks in Becerra-Suarez et al. [37] and Multi-Layer Perceptron in Siva Shankar et al. [33]. This is a direct consequence of the increased model complexity of DL solutions, which makes them perform better compared to standard ML techniques, especially in classification tasks. The employment of AI is mainly aimed at the identification and classification of cyber threats, especially for what concerns Intrusion Detection Systems, given the predominance of intrusion-based attacks in the analyzed works, with malware and DoS being the most common cyberattacks identified in the analyzed works.

However, some common criticalities emerge that limit the adoption and generalization of these approaches in real-world contexts. A recurring aspect concerns the dependence on obsolete or highly simplified datasets, such as the KDD Cup 99 and the NSL-KDD, which are used in several included works (e.g., Geetha et al. [29], Siva Shankar et al. [33] and Abdullahi et al. [12]), with only a few works such as Aldhyani et al. [39] and Obonna et al. [41] using datasets constructed from real data. While useful for comparable assessments, such datasets fail to represent the complexity and dynamics of real cyber environments, especially in the presence of new types of attacks (e.g., zero-day attacks [48]) and in emerging contexts such as IoT and smart grids. Moreover, many new attacks tailored directly for AI should also be included in new datasets, such as model collapse and gradient leakage [49]. The presence of unbalanced data and the scarcity of samples for some attack classes represents a further limitation that can compromise the generalization of models [50], as confirmed by works such as Bhandari et al. [35], which also identifies potential overfitting issues due to class imbalance. This points to the necessity to work on new, balanced datasets which can make the already explored models more robust to lesser-known attacks, either via re-training of the already trained data or by performing fine tuning [51] or transfer learning [52,53] on the new datasets.

Another shared limitation concerns the high computational requirement, in particular for DL-based architectures. Many studies (such as Gaba et al. [40] and Geetha et al. [29]) report that, although models achieve excellent accuracy in the training and validation

phase on benchmark datasets, their real-time use (especially in environments with limited resources, such as IoT devices or in distributed systems) requires hardware infrastructures that are not always available or involve high implementation costs. This is typically related to the high number of model parameters of more complex architectures such as deep neural networks, which typically results in a high number of Floating Point Operations Per Second (FLOPS) and memory usage at training and inference time. For this reason, several studies are moving either towards the reduction of FLOPS [54] or through an increase in speed by making FLOPS more efficient [55].

The aforementioned problem is correlated to, and highly coincident with, a fundamental gap in AI, scalability, which is often acknowledged in the literature. Scalability is a problem which is often mentioned, with Mishra et al. [18] even providing information about the increase in scalability with the application of an EES + KNN protocol compared with standard ML techniques (showing an increase in scalability of up to 18%). However, usually authors, such as Geetha et al. [29], Bendiab et al. [38] and Gaba et al. [40], note scalability problems in currently implemented algorithms, especially for what concerns real-time scalability. While scalability is more tied to performance degradation in the case of big amounts of data, like is the case in real-time applications, this problem is mostly coincident with the computational requirement of models. Consequently, the reduction in model complexity also partially solves the problem of model scalability. However, some dedicated solutions are also emerging to improve model scalability, such as serverless computing, especially for what concerns model inference [56].

A related issue is the low interpretability of many solutions based on DL models. While some argue that the "black box" phenomenon could have its advantages [57], it undoubtedly limits operators' ability to understand the reasons behind model decisions, a critical aspect in the context of cybersecurity where transparency and the ability to explain any errors are essential for trust and accountability management. This motivates the recent increase in interest in xAI solutions [58], which can be fundamental in explaining decisional mechanisms of models and, consequently, to understand where in the models errors may occur and possibly how to resolve them [59]. Moreover, seeing how some authors noted the rise of ML-directed attacks in recent years (such as Wazid et al. [21], Yu et al. [23] and Teo et al. [44]), gaining a better understanding of the algorithms could also help in the identification of criticalities in models which might be subject to cyberattacks, allowing us to preemptively strengthen them accordingly instead of focusing on further defense mechanisms.

At the same time, transparency is closely tied to low interpretability, as a non-interpretable model consequently has low transparency. Several authors specifically note the limitations on model transparency, such as Polito et al. [20], Bendiab et al. [38] and Pawlicki et al. [45], with all of them citing as a solution the introduction of xAI to understand the behaviors of models. For this reason, investing in xAI research is particularly relevant in order to generate increasingly transparent models which can easily be analyzed and strengthened with more tailored cybersecurity solutions.

Similarly, it is important to highlight the risk associated with adversarial attacks and data poisoning, which can compromise the robustness of the system despite the high declared performance. The rapidly evolving threat landscape also requires continuous updating of algorithms and the integration of human supervision systems (for example, with the adoption of "kill switches", which, however, introduce some vulnerabilities [60]) to ensure that the solutions developed can effectively adapt to ever-changing scenarios. While Pawlicki et al. [45] and Lee et al. [46] mention the possibility of working towards adaptive systems through the use of xAI, such steps have not yet been taken. Moreover, while xAI has proven to be a strong tool against adversarial attacks [61], xAI itself can also

be employed to enforce adversarial attacks [62], making it an equally useful and dangerous tool. What emerges, therefore, is that constant human supervision is required in order to ensure the quality of employed algorithms, which can be easily subjected to decreases in performance in the case of external malicious intervention.

Surprisingly, quantum-safe techniques are rarely acknowledged, with some of the only exceptions being Radanliev et al. [22] and Bendiab et al. [38]. In a landscape where quantum computers are quickly becoming more influential and with an expected rapidly decreasing cost, it is particularly relevant to focus on solutions to prevent attacks from such devices. Some emerging solutions in this direction are Post-Quantum Cryptography (POC), which has the goal of introducing new cryptography algorithms such as lattice-based, hash-based, and code-based cryptography to reduce the risk of attacks from quantum computers [63], and Quantum Key Distribution (QKD), which instead aims at identifying eavesdropping attempts with quantum mechanics and altering the quantum state of the key if a potential attack has been detected [64]. However, applications of AI to this field are rarely found in the literature, and while it is acknowledged to be a promising direction, still lacks scalability protocols and sufficient security [65].

In summary, while on one hand there is great potential in the use of AI to enhance cybersecurity, which has been evidenced by the selected literature in this review given the excellent analytical results obtained on several different tasks such as intrusion detection and threat identification, on the other there are still limitations and open questions related to data quality, high computational demands, the difficulty in interpreting decisions, and the need to ensure robustness against emerging threats. These critical issues suggest the importance of developing hybrid and multidisciplinary approaches, integrating interpretable ML solutions and human control systems, as well as the need to continuously update datasets and methodologies to more accurately reflect the real dynamics of cyber environments.

4.1. Future Works

Following from the previously highlighted limitations, several directions for future developments can be defined. First of all, as previously mentioned, most of the analyzed studies are based on traditional datasets (such as NSL-KDD, KDD Cup 99, or similar benchmarks) that, although useful for comparative analyses, do not fully reflect the complexity and dynamism of real-world scenarios. Therefore, a key direction for future research consists in the development and adoption of up-to-date datasets representative of the current dynamics of cyberspace, especially in emerging environments such as the Internet of Things, smart grids and cyber-physical systems. Collecting data from operational environments, with particular attention to class balance and bias management, would allow training more robust and generalizable models, as Aldhyani et al. [39] and Obonna et al. [41] demonstrate. Alternatively, in the impossibility of creating datasets based on real environments (for example, complex environments or limited quantity of available data), generative AI could also be employed for the generation of new data samples, a solution already tested in the medical field [66]. As data created by generative AI is subject to errors and to the same adversarial attacks commonly evidenced in previously discussed solutions, however, the preferred direction would still be that of real data collection. Finally, there are also solutions in the literature for limited-size datasets [67], which could also be a possible direction in the cybersecurity field.

Another area of urgent improvement concerns the interpretability of DL-based models. Due to many of the described algorithms operating as "black boxes" (like Halbouni et al. [31], Pawlicki et al. [45] and Lee at al. [46] highlight), future research should deepen and integrate xAI methodologies, developing techniques that combine the predictive power of models with interpretation tools, for example through the use of methods such as LIME,

SHAP, rule-based or explanatory decision trees [45]. Such an approach would not only increase operator confidence, but also facilitate error management and response, a crucial element in security applications. By increasing human supervision of algorithms, it is possible to intervene in several evidenced critical areas, such as the identification of data poisoning and adversarial attacks, or tackling overfitting and the effects of class imbalance. Finally, xAI can also help in the creation of adaptive system which can automatically reconfigure themselves to new, unseen threats without needing human intervention, which would make attack response faster and consequently improve security. For this reason, investing in research in xAI is particularly relevant, and should be one of the main directions to follow.

In parallel, there is a need to address the problem of the high computational demand typical of DL architectures. The adoption of real-time AI models, especially in resource-constrained environments such as IoT devices, requires the implementation of lightweight and distributed solutions. Techniques such as federated learning (as suggested by Bendiab et al. [38], Shees et al. [42] and Teo et al. [44]), few-shot learning (as highlighted by Abdullahi et al. [12], and hybrid approaches combining traditional methods and advanced algorithms (like many authors, such as Al-Quayed et al. [24] and Pawlicki et al. [45], suggest) could help reduce the computational load, while ensuring a timely and accurate response to threats. Alternatively, current architectures should be reworked by reducing the number of operations while keeping constant, or possibly increasing, model accuracy. A solution could be introducing novel layers, which can dynamically adjust kernel size, a solution which has already been tested in other domains such as crowd counting [68]. Furthermore, it is necessary to develop actionable and dynamic defense strategies to counter adversarial attacks and data poisoning phenomena besides human supervision, thus ensuring greater resilience of AI systems even in complex operational scenarios.

Finally, the described state of the art, as shown by articles such as Radanliev et al. [22], suggests the importance of a multidisciplinary approach, which combines technical, regulatory, and risk management skills. The integration between AI systems, human supervision (for example through the implementation of "kill switches", like Polito et al. [20] suggest, or override mechanisms), and the harmonization of international standards for cybersecurity (such as ISO/IEC, NIST, or ISA/IEC 62443, as authors such as Al-Quayed et al. [24], Teo et al. [44] and Radanliev et al. [22] highlight) represents a strategic way to address emerging threats and foster cyber diplomacy. A close collaboration between researchers, security professionals, and public policy makers could facilitate the development of integrated security frameworks, capable of dynamically adapting to the evolutions of the digital context. In such a context, regulatory systems should aim towards unification, as they remain mostly heterogeneous both in terms of applications and geographically.

In short, future research should not only focus on algorithm optimization and dataset updating, but also on investing in transparent, scalable and cyber-attack-resistant AI models.

4.2. Limitations of the Study

While our study provided a comprehensive analysis of the available literature through a systematic review process guided by the PRISMA protocol, it is still subject to limitations. The main one comes from the absence in our study of an analysis of gray literature, which has not been analyzed. Therefore, relevant studies that are not indexed online may not have been included in our study, potentially limiting the extent of our review. Similarly, works not in English and works more recent than April 2025 have been excluded from our study, further reducing the scope of considered articles. Finally, our research has been conducted

only through the Google Scholar database, which could have potentially excluded works from other relevant databases such as Scopus and PubMed.

Due to the subjective nature of the screening process, which while guided by inclusion and exclusion criteria was a manual process, it is possible that reviewer bias favored some works compared to others, rendering our literature selection imprecise or subject to selection bias. Moreover, the results of the included studies were not reproduced, which makes our analysis only theoretical and based on the results reported by the authors of the included studies. This causes some limitation in the comparison of metrics, as the majority of works have been tested on extensively different datasets and domains and are therefore only partially comparable. Finally, no statistic analysis has been performed on the obtained results, making it impossible to verify codependency of effects or correlations between the obtained results.

Our work also limited the scope of applications to only certain domains (industry, infrastructure, finance). Future reviews could focus on a more complete cross-sectoral analysis by analyzing if and how AI methods for cybersecurity are applicable to vastly different domains with different security standards. A statistical analysis was also not performed for two main reasons. The first is the scoping review framework of our research, which is more focused on mapping literature and identifying gaps rather than performing a qualitative synthesis, as confirmed by the PRISMA-ScR checklist. The second is the heterogeneity of the included works, which makes a uniform comparison unfeasible or heavily subjected to bias. Future reviews should therefore adopt a systematic review framework or perform a meta-analysis on a homogeneous subset of the results (e.g., by selecting a single domain of application) in order to confirm the results found in current literature.

5. Conclusions

In this review, we explored the use of AI in cybersecurity, focusing on critical and emerging sectors such as Industry 4.0, IoT, Blockchain, Finance and Smart Grids. The analysis involved 24 studies selected with rigorous criteria, which compared various ML and DL algorithms, including Support Vector Machine, Random Forest, Decision Trees, K-Nearest Neighbors, CNN, RNN, LSTM, Autoencoder, and hybrid approaches integrated with fuzzy logic, evaluating their applicability in intrusion detection, endpoint protection, anti-phishing and network resilience.

The overall results highlight promising performances, with frequently high accuracies (often over 95–99% on classic benchmarks such as NSL-KDD, KDD Cup 99, UNSW-IDS15, and CICIDS2017) and a significant ability to identify anomalies and modulate responses to attacks. The most commonly used algorithms are DT and RF, while the best results are obtained by CNN and MLP. The most common use of AI in cybersecurity is for IDS and attack classification, with the most frequent cyberattacks identified being malware and DoS. The excellent analytical results found in the literature confirm that the answer to our research question, 'Is AI an effective method to enhance current infrastructures' cybersecurity?', is affirmative. However, significant critical issues have emerged related to the use of sometimes obsolete or highly simplified datasets, the high computational requirement of DL models, and the poor interpretability of "black box" solutions. These aspects limit the generalizability of the solutions, especially in real operational scenarios characterized by complex dynamics and limited computational resources.

The future lines of research that emerge from this study are crucial to overcome these limitations, and are highlighted in Figure 3. First, the development and adoption of updated datasets representative of the real dynamics of cyberspace is essential, especially in emerging environments such as IoT, smart grids, and cyber-physical systems. These datasets should also ensure better management of imbalances and biases, thus allowing

more robust and generalizable training of the models, while also including more recent cyberattacks that have emerged in a new landscape of threats such as identity theft. In parallel, it is necessary to invest in the development of xAI techniques, employing methods such as LIME, SHAP, and explanatory rule-based approaches to make model decision-making processes transparent and easily interpretable, strengthening operator trust and facilitating error management. Currently, many algorithms operate as a 'black-box', while clear interpretability of the models will greatly improve the quality of human intervention in current frameworks by highlighting which sections of the model and which parameters have the most impact on the results and where to intervene to improve it or strengthen it.

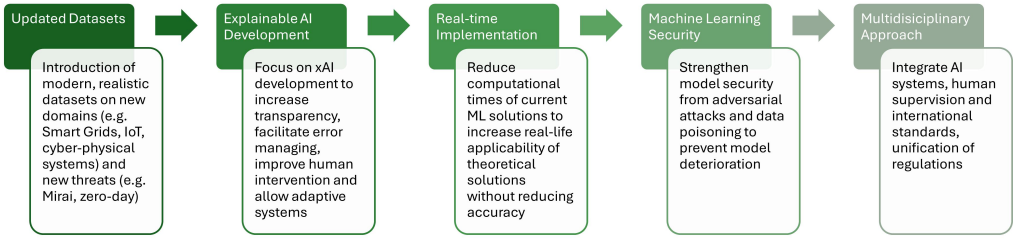

Figure 3. Timeline for future research in AI for cybersecurity.

Another aspect concerns the need to optimize models for real-time implementation, especially in contexts with limited resources. The application of lightweight solutions, such as federated learning, few-shot learning, and hybrid approaches that combine traditional techniques and advanced algorithms, can help reduce the computational load without compromising accuracy. The reduction of computational latency can greatly strengthen the security of systems by allowing immediate intervention against cyber threats, and the aforementioned techniques can work in this direction by promoting collaborative parallel learning, thus reducing the individual computational load (federated learning), or by reducing training samples to make the model faster (few-shot learning). At the same time, it is essential to develop defense strategies against adversarial attacks and data poisoning phenomena, thus ensuring the robustness and reliability of AI systems even in complex and dynamic scenarios. This is particularly relevant as direct attacks to AI models used to enforce security might generate openings for further attacks or heavily compromise the protection of cyber systems. At the same time, a special focus should be placed on quantum-safe techniques such as POC and QKD, as the rise in quantum computing could easily disrupt currently existing techniques.

Finally, the cybersecurity landscape requires a multidisciplinary approach that integrates technical, regulatory, and risk management skills. The fundamental step to follow should be the integration of AI systems, human supervision, and the harmonization of international standards while guaranteeing a unification of the already existing regulations to strengthen cyber diplomacy. Examples of the integration between AI and humans would be with the introduction of "kill switches" or override mechanisms.

In summary, while on the one hand the use of AI in cybersecurity offers clear advantages in terms of precise detection and rapid response to attacks, on the other hand, there is still the need to overcome critical issues related to datasets, computational efficiency, and model interpretability. The future directions outlined in this work provide concrete ideas to further enhance the effectiveness of AI-based solutions, ensuring more scalable, transparent, and resilient systems for the evolution of cyberspace.

Author Contributions: Conceptualization, C.R. and C.N.; methodology, C.R. and C.N.; software, F.F. and K.L.; validation, F.F. and C.R.; formal analysis, C.N.; investigation, C.R., F.F. and K.L.; resources, C.N.; data curation, K.L.; writing—original draft preparation, C.R., F.F., K.L. and C.N.; writing—review and editing, F.F., C.R. and C.N.; supervision, C.N.; project administration, C.N.; funding acquisition, C.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The datasets utilized in this study are publicly available.

Acknowledgments: This work has been developed at *is.Lab()* Intelligent Systems Laboratory of the Department of Computer, Control, and Management Engineering, Sapienza University of Rome. This work has been carried out while Francesca Fiani was enrolled in the Italian National Doctorate on Artificial Intelligence run by Sapienza University of Rome in collaboration with the Department of Computer, Control, and Management Engineering.

Conflicts of Interest: The authors declare no potential conflicts of interest.

References

1. Kemmerer, R.A. Cybersecurity. In Proceedings of the 25th International Conference on Software Engineering, Portland, OR, USA, 3–10 May 2003; Proceedings; IEEE: New York, NY, USA, 2003; pp. 705–715.
2. Boiko, A.; Shendryk, V.; Boiko, O. Information systems for supply chain management: Uncertainties, risks and cyber security. *Procedia Comput. Sci.* **2019**, *149*, 65–70. [CrossRef]
3. Coventry, L.; Branley, D. Cybersecurity in healthcare: A narrative review of trends, threats and ways forward. *Maturitas* **2018**, *113*, 48–52. [CrossRef] [PubMed]
4. Uddin, M.H.; Ali, M.H.; Hassan, M.K. Cybersecurity hazards and financial system vulnerability: A synthesis of literature. *Risk Manag.* **2020**, *22*, 239–309. [CrossRef]
5. Ten, C.W.; Manimaran, G.; Liu, C.C. Cybersecurity for critical infrastructures: Attack and defense modeling. *IEEE Trans. Syst. Man-Cybern.-Part Syst. Humans* **2010**, *40*, 853–865. [CrossRef]
6. Lezzi, M.; Lazoi, M.; Corallo, A. Cybersecurity for Industry 4.0 in the current literature: A reference framework. *Comput. Ind.* **2018**, *103*, 97–110. [CrossRef]
7. Di Pierro, M. What is the blockchain? *Comput. Sci. Eng.* **2017**, *19*, 92–95. [CrossRef]
8. Moreno Escobar, J.J.; Morales Matamoros, O.; Tejeida Padilla, R.; Lina Reyes, I.; Quintana Espinosa, H. A comprehensive review on smart grids: Challenges and opportunities. *Sensors* **2021**, *21*, 6978. [CrossRef]
9. Mohamed, N. Current trends in AI and ML for cybersecurity: A state-of-the-art survey. *Cogent Eng.* **2023**, *10*, 2272358. [CrossRef]
10. Admass, W.S.; Munaye, Y.Y.; Diro, A.A. Cyber security: State of the art, challenges and future directions. *Cyber Secur. Appl.* **2024**, *2*, 100031. [CrossRef]
11. Zhang, Z.; Ning, H.; Shi, F.; Farha, F.; Xu, Y.; Xu, J.; Zhang, F.; Choo, K.K.R. Artificial intelligence in cyber security: Research advances, challenges, and opportunities. *Artif. Intell. Rev.* **2022**, *55*, 1029–1053. [CrossRef]
12. Abdullahi, M.; Baashar, Y.; Alhussian, H.; Alwadain, A.; Aziz, N.; Capretz, L.F.; Abdulkadir, S.J. Detecting cybersecurity attacks in internet of things using artificial intelligence methods: A systematic literature review. *Electronics* **2022**, *11*, 198. [CrossRef]
13. Handa, A.; Sharma, A.; Shukla, S.K. Machine learning in cybersecurity: A review. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2019**, *9*, e1306. [CrossRef]
14. Abbas, N.N.; Ahmed, T.; Shah, S.H.U.; Omar, M.; Park, H.W. Investigating the applications of artificial intelligence in cyber security. *Scientometrics* **2019**, *121*, 1189–1211. [CrossRef]
15. Apruzzese, G.; Laskov, P.; Montes de Oca, E.; Mallouli, W.; Brdalo Rapa, L.; Grammatopoulos, A.V.; Di Franco, F. The role of machine learning in cybersecurity. *Digit. Threat. Res. Pract.* **2023**, *4*, 1–38. [CrossRef]
16. Randieri, C.; Perrotta, A.; Puglisi, A.; Grazia Bocci, M.; Napoli, C. CNN-Based Framework for Classifying COVID-19, Pneumonia, and Normal Chest X-Rays. *Big Data Cogn. Comput.* **2025**, *9*, 186. [CrossRef]
17. Gunning, D.; Stefik, M.; Choi, J.; Miller, T.; Stumpf, S.; Yang, G.Z. XAI—Explainable artificial intelligence. *Sci. Robot.* **2019**, *4*, eaay7120. [CrossRef]
18. Mishra, S. Exploring the impact of AI-based cyber security financial sector management. *Appl. Sci.* **2023**, *13*, 5875. [CrossRef]

19. Llorca, D.F.; Hamon, R.; Junklewitz, H.; Grosse, K.; Kunze, L.; Seiniger, P.; Swaim, R.; Reed, N.; Alahi, A.; Gómez, E.; et al. Testing autonomous vehicles and AI: Perspectives and challenges from cybersecurity, transparency, robustness and fairness. *Eur. Transp. Res. Rev.* **2025**, *17*, 38. [CrossRef]
20. Polito, C.; Pupillo, L. Artificial intelligence and cybersecurity. *Intereconomics* **2024**, *59*, 10–13. [CrossRef]
21. Wazid, M.; Das, A.K.; Chamola, V.; Park, Y. Uniting cyber security and machine learning: Advantages, challenges and future research. *ICT Express* **2022**, *8*, 313–321. [CrossRef]
22. Radanliev, P. Cyber diplomacy: Defining the opportunities for cybersecurity and risks from Artificial Intelligence, IoT, Blockchains, and Quantum Computing. *J. Cyber Secur. Technol.* **2025**, *9*, 28–78. [CrossRef]
23. Yu, J.; Shvetsov, A.V.; Alsamhi, S.H. Leveraging machine learning for cybersecurity resilience in industry 4.0: Challenges and future directions. *IEEE Access* **2024**, *12*, 159579–159596. [CrossRef]
24. Al-Quayed, F.; Ahmad, Z.; Humayun, M. A situation based predictive approach for cybersecurity intrusion detection and prevention using machine learning and deep learning algorithms in wireless sensor networks of industry 4.0. *IEEE Access* **2024**, *12*, 34800–34819. [CrossRef]
25. *ISO/IEC 27001:2022*; Information Security, Cybersecurity and Privacy Protection—Information Security Management Systems—Requirements. International Organization for Standardization (ISO); International Electrotechnical Commission (IEC): Geneva, Switzerland, 2022.
26. National Institute of Standards and Technology. *Security Requirements for Cryptographic Modules*; Department of Commerce, Federal Information Processing Standards Publications (FIPS) 140-2: Washington, DC, USA, 2002.
27. *ISA-62443-1-1-2007*; Security for Industrial Automation and Control Systems, Part 1-1: Terminology, Concepts, and Models. International Society of Automation: Durham, NC, USA, 2007.
28. Deshpande, A. Cybersecurity in Financial Services: Addressing AI-Related Threats and Vulnerabilities. In Proceedings of the 2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS), Chikkaballapur, India, 18–19 April 2024; IEEE: New York, NY, USA, 2024; Volume 1, pp. 1–6.
29. Geetha, R.; Thilagam, T. A review on the effectiveness of machine learning and deep learning algorithms for cyber security. *Arch. Comput. Methods Eng.* **2021**, *28*, 2861–2879. [CrossRef]
30. Boutaba, R.; Salahuddin, M.A.; Limam, N.; Ayoubi, S.; Shahriar, N.; Estrada-Solano, F.; Caicedo, O.M. A comprehensive survey on machine learning for networking: Evolution, applications and research opportunities. *J. Internet Serv. Appl.* **2018**, *9*, 16. [CrossRef]
31. Halbouni, A.; Gunawan, T.S.; Habaebi, M.H.; Halbouni, M.; Kartiwi, M.; Ahmad, R. Machine learning and deep learning approaches for cybersecurity: A review. *IEEE Access* **2022**, *10*, 19572–19585. [CrossRef]
32. Dell'Olmo, P.V.; Kuznetsov, O.; Frontoni, E.; Arnesano, M.; Napoli, C.; Randieri, C. Dataset Dependency in CNN-Based Copy-Move Forgery Detection: A Multi-Dataset Comparative Analysis. *Mach. Learn. Knowl. Extr.* **2025**, *7*, 54. [CrossRef]
33. Siva Shankar, S.; Hung, B.T.; Chakrabarti, P.; Chakrabarti, T.; Parasa, G. A novel optimization based deep learning with artificial intelligence approach to detect intrusion attack in network system. *Educ. Inf. Technol.* **2024**, *29*, 3859–3883. [CrossRef]
34. Nabi, F.; Zhou, X. Enhancing intrusion detection systems through dimensionality reduction: A comparative study of machine learning techniques for cyber security. *Cyber Secur. Appl.* **2024**, *2*, 100033. [CrossRef]
35. Bhandari, G.; Lyth, A.; Shalaginov, A.; Grønli, T.M. Distributed deep neural-network-based middleware for cyber-attacks detection in smart IoT ecosystem: A novel framework and performance evaluation approach. *Electronics* **2023**, *12*, 298. [CrossRef]
36. Priya, S.S.; Sanjana, P.S.; Yanamala, R.M.R.; Amar Raj, R.D.; Pallakonda, A.; Napoli, C.; Randieri, C. Flight-Safe Inference: SVD-Compressed LSTM Acceleration for Real-Time UAV Engine Monitoring Using Custom FPGA Hardware Architecture. *Drones* **2025**, *9*, 494. [CrossRef]
37. Becerra-Suarez, F.L.; Tuesta-Monteza, V.A.; Mejia-Cabrera, H.I.; Arcila-Diaz, J. Performance evaluation of deep learning models for classifying cybersecurity attacks in IoT networks. *Informatics* **2024**, *11*, 32. [CrossRef]
38. Bendiab, G.; Hameurlaine, A.; Germanos, G.; Kolokotronis, N.; Shiaeles, S. Autonomous vehicles security: Challenges and solutions using blockchain and artificial intelligence. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 3614–3637. [CrossRef]
39. Aldhyani, T.H.; Alkahtani, H. Attacks to automatous vehicles: A deep learning algorithm for cybersecurity. *Sensors* **2022**, *22*, 360. [CrossRef]
40. Gaba, S.; Budhiraja, I.; Kumar, V.; Martha, S.; Khurmi, J.; Singh, A.; Singh, K.K.; Askar, S.S.; Abouhawwash, M. A systematic analysis of enhancing cyber security using deep learning for cyber physical systems. *IEEE Access* **2024**, *12*, 6017–6035. [CrossRef]
41. Obonna, U.O.; Opara, F.K.; Mbaocha, C.C.; Obichere, J.K.C.; Akwukwaegbu, I.O.; Amaefule, M.M.; Nwakanma, C.I. Detection of Man-in-the-middle (MitM) cyber-attacks in oil and gas process control networks using machine learning algorithms. *Future Internet* **2023**, *15*, 280. [CrossRef]
42. Shees, A.; Tariq, M.; Sarwat, A.I. Cybersecurity in Smart Grids: Detecting False Data Injection Attacks Utilizing Supervised Machine Learning Techniques. *Energies* **2024**, *17*, 5870. [CrossRef]

43. Shaukat, K.; Luo, S.; Varadharajan, V.; Hameed, I.A.; Chen, S.; Liu, D.; Li, J. Performance comparison and current challenges of using machine learning techniques in cybersecurity. *Energies* **2020**, *13*, 2509. [CrossRef]
44. Teo, Z.L.; Ning, C.Q.W.; Wong, J.L.Y.; Ting, D. Cybersecurity in the generative artificial intelligence era. *Asia-Pac. J. Ophthalmol.* **2024**, *13*, 100091. [CrossRef]
45. Pawlicki, M.; Pawlicka, A.; Kozik, R.; Choraś, M. Advanced insights through systematic analysis: Mapping future research directions and opportunities for xAI in deep learning and artificial intelligence used in cybersecurity. *Neurocomputing* **2024**, *590*, 127759. [CrossRef]
46. Lee, H.W.; Han, T.H.; Lee, T.J. Reference-Based AI Decision Support for Cybersecurity. *IEEE Access* **2023**, *11*, 143324–143339. [CrossRef]
47. Napoli, C.; Napoli, C.; Ponzi, V.; Puglisi, A.; Russo, S.; Tibermacine, I.E. Exploiting Robots as Healthcare Resources for Epidemics Management and Support Caregivers. In Proceedings of the 10th Italian Workshop on Artificial Intelligence and Robotics, Rome Italy, 9 November 2023; Volume 3686, pp. 1–10.
48. Ahmad, R.; Alsmadi, I.; Alhamdani, W.; Tawalbeh, L. Zero-day attack detection: A systematic literature review. *Artif. Intell. Rev.* **2023**, *56*, 10733–10811. [CrossRef]
49. De Maio, C.; Di Gisi, M.; Fenza, G.; Gallo, M.; Loia, V. A Lifecycle-Oriented Survey of Emerging Threats and Vulnerabilities in Large Language Models. *IEEE Access* **2025**, *13*, 176482–176500. [CrossRef]
50. Althnian, A.; AlSaeed, D.; Al-Baity, H.; Samha, A.; Dris, A.B.; Alzakari, N.; Abou Elwafa, A.; Kurdi, H. Impact of dataset size on classification performance: An empirical evaluation in the medical domain. *Appl. Sci.* **2021**, *11*, 796. [CrossRef]
51. Wang, J.Q.; Guo, L.; Jiang, Y.; Zhang, S.; Zhou, Q. Improving unbalanced image classification through fine-tuning method of reinforcement learning. *Appl. Soft Comput.* **2024**, *163*, 111841. [CrossRef]
52. Soekhoe, D.; Van Der Putten, P.; Plaat, A. On the impact of data set size in transfer learning using deep neural networks. In Proceedings of the International Symposium on Intelligent Data Analysis, Stockholm, Sweden, 13–15 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 50–60.
53. Milicevic, M.; Zubrinic, K.; Obradovic, I.; Sjekavica, T. Data augmentation and transfer learning for limited dataset ship classification. *WSEAS Trans. Syst. Control* **2018**, *13*, 460–465.
54. Hsia, S.C.; Wang, S.H.; Chang, C.Y. Convolution neural network with low operation FLOPS and high accuracy for image recognition. *J. Real-Time Image Process.* **2021**, *18*, 1309–1319. [CrossRef]
55. Chen, J.; Kao, S.H.; He, H.; Zhuo, W.; Wen, S.; Lee, C.H.; Chan, S.H.G. Run, don't walk: Chasing higher FLOPS for faster neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 12021–12031.
56. Wang, L.; Jiang, Y.; Mi, N. Advancing serverless computing for scalable ai model inference: Challenges and opportunities. In Proceedings of the 10th International Workshop on Serverless Computing, Hong Kong, China, 2–6 December 2024; pp. 1–6.
57. Loyola-Gonzalez, O. Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE Access* **2019**, *7*, 154096–154113. [CrossRef]
58. Capuano, N.; Fenza, G.; Loia, V.; Stanzione, C. Explainable artificial intelligence in cybersecurity: A survey. *IEEE Access* **2022**, *10*, 93575–93600. [CrossRef]
59. Pahde, F.; Dreyer, M.; Samek, W.; Lapuschkin, S. Reveal to revise: An explainable ai life cycle for iterative bias correction of deep models. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Vancouver, BC, Canada, 8–12 October 2023; Springer: Berlin/Heidelberg, Germany, 2023; pp. 596–606.
60. Oravec, J.A. Kill switches, remote deletion, and intelligent agents: Framing everyday household cybersecurity in the internet of things. *Technol. Soc.* **2017**, *51*, 189–198. [CrossRef]
61. Fenza, G.; Loia, V.; Stanzione, C.; Di Gisi, M. Robustness of models addressing Information Disorder: A comprehensive review and benchmarking study. *Neurocomputing* **2024**, *596*, 127951. [CrossRef]
62. Cavaliere, D.; Gallo, M.; Stanzione, C. Propaganda detection robustness through adversarial attacks driven by explainable ai. In Proceedings of the World Conference on Explainable Artificial Intelligence, Lisbon, Portugal, 26–28 July 2023; Springer: Berlin/Heidelberg, Germany, 2023; pp. 405–419.
63. Dam, D.T.; Tran, T.H.; Hoang, V.P.; Pham, C.K.; Hoang, T.T. A survey of post-quantum cryptography: Start of a new race. *Cryptography* **2023**, *7*, 40. [CrossRef]
64. Cao, Y.; Zhao, Y.; Wang, Q.; Zhang, J.; Ng, S.X.; Hanzo, L. The evolution of quantum key distribution networks: On the road to the qinternet. *IEEE Commun. Surv. Tutorials* **2022**, *24*, 839–894. [CrossRef]
65. Radanliev, P. Artificial intelligence and quantum cryptography. *J. Anal. Sci. Technol.* **2024**, *15*, 4. [CrossRef]
66. Umer, F.; Adnan, N. Generative artificial intelligence: Synthetic datasets in dentistry. *BDJ Open* **2024**, *10*, 13. [CrossRef]

67. Shaikhina, T.; Khovanova, N.A. Handling limited datasets with neural networks in medical applications: A small-data approach. *Artif. Intell. Med.* **2017**, *75*, 51–63. [CrossRef]
68. Tomar, A.; Kumar, S.; Pant, B.; Tiwari, U.K. Dynamic Kernel CNN-LR model for people counting. *Appl. Intell.* **2022**, *52*, 55–70. [CrossRef]