

AI in Cybersecurity: Ethical Challenges of AI-Based Intrusion Detection Systems: A Comparative Analysis of EU, U.S., and South Korea

Rihab Karoud^{1*}

¹Network Security Department, Higher Institute of Information and Communication Technologies, University of Carthage, Tunisia

*Corresponding Author: rihabkaroud1@gmail.com

Article info: Received March 22, 2026, Revised May 04, 2026, Accepted May 9, 2026

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Abstract

The concept of artificial intelligence (AI) has been increasingly used in the field of cybersecurity, especially in the development of intrusion detection systems (IDS). The conventional approach to developing intrusion detection systems involves the use of rule-based systems as well as signatures of attacks. However, this approach has been criticized for its inability to deal with emerging threats such as zero-day exploits. This has led to the development of intrusion detection systems that incorporate the concept of machine learning in the analysis of network activities. This approach has been effective in the detection of anomalies that could be indicative of malicious activities. However, the use of this approach in the development of intrusion detection systems has raised ethical issues. This paper seeks to discuss the ethical issues surrounding the application of AI-based intrusion detection systems by first presenting an overview of the technical basis of the systems and the technical benefits they offer. Secondly, the paper will discuss the ethical issues surrounding the application of the system. Some of the ethical issues discussed will include the lack of explainability of machine learning algorithms, the possibility of biased decision-making by the system, and the ethical issues surrounding the collection of large amounts of personal information. Additionally, the paper will discuss the technical solutions to the ethical issues surrounding the application of the system. Some of the technical solutions to the ethical issues will include the application of explainable AI, federated learning, differential privacy, and synthetic data. Lastly, the paper will discuss the regulatory frameworks adopted by the European Union, the US, and South Korea regarding the application of the system. The analysis identifies the gaps in the regulations as well as the technical limitations in the ethical use of AI-based IDS. The study offers recommendations to improve the design of AI-based IDS. The study emphasizes the need to strike a balance between the increasing technical capabilities of AI systems and the need to have proper ethical and regulatory checks in the use of AI in cybersecurity.

Keywords: cybersecurity, Artificial intelligence, intrusion detection systems, ethical challenges, network activities

1. Introduction

Due to the rapid digitalization of modern societies, the role of cybersecurity has gained considerable significance. Organizations across various industries are heavily relying on interconnected systems, cloud services, and large-scale data infrastructures to support day-to-day operations [1]. Although these emerging technologies are helping to improve the efficiency and innovation of the systems, they are also making the systems more vulnerable to cyber-attacks. Intrusion Detection Systems (IDS) play an important role in identifying malicious activities within computer networks and computer systems. Traditional Intrusion Detection Systems' techniques mainly depend on signature-based or rule-based detection techniques. These techniques can successfully identify malicious attacks on computer networks and computer systems. However, traditional techniques may not be effective in

identifying new attacks or malicious activities on computer networks and computer systems [2]. In addition, the complexity and volume of network traffic make traditional detection techniques less effective and make manual detection more difficult to implement.

In order to overcome the limitations of the above-mentioned methods, artificial intelligence and machine learning techniques have been widely used in modern cybersecurity systems. AI-based intrusion detection systems can analyze the large amount of network traffic data and learn the patterns of normal behavior to detect any anomalous activities [2]. Moreover, the potential of deep learning techniques for detecting unknown types of attacks by automatically learning the relevant features from the large amount of network traffic data has been demonstrated. [12] – [14]. Despite all the above technical benefits, the use of AI in intrusion detection systems poses new issues of ethics and governance. According to [23], machine learning systems tend to operate as black box systems. It is difficult to understand the decision made by the system. Moreover, the training data used in machine learning systems may have bias in their composition [50]. There is a problem of privacy as well. This is due to the fact that intrusion detection systems often require the monitoring of all the activities in the network [57].

This paper seeks to investigate the technical basis and ethical concerns surrounding AI-based intrusion detection systems. It focuses on the application of AI techniques like machine learning and deep learning in modern IDS systems and discusses the ethical concerns surrounding AI-based intrusion detection systems. In addition, the paper offers a comparative analysis of the regulatory environment in the European Union, the United States, and South Korea to identify differences and similarities in their governance and operation. This paper hopes to make a contribution to the field of AI-based intrusion detection systems by identifying gaps and making recommendations to improve transparency and ethical concerns surrounding AI-based intrusion detection systems. To guide the analysis, the study addresses the following research questions:

RQ1: What are the main ethical challenges associated with the deployment of artificial intelligence in intrusion detection systems?

RQ2: How do regulatory frameworks in the European Union, the United States, and South Korea address issues of transparency, privacy, and governance in AI-based cybersecurity systems?

RQ3: What technical and policy approaches can help mitigate ethical risks and support the responsible deployment of AI-driven intrusion detection systems?

2. Research Methodology

In conducting this research, a qualitative research approach will be used which will comprise an analysis of structured literature review, and a comparative analysis approach will be used. This will be done to understand the ethical issues related to AI-based IDS and the manner in which various regulatory frameworks tackle them. Sources of secondary information include academic journals, conference papers, and regulatory documents such as GDPR, the US NIST framework, and PIPA in South Korea. The selection of the sources was made considering that the sources deal with artificial intelligence and cyber-security, ethics, and IDS. Comparative analysis is a method which can be used to compare the approaches taken by various states including the EU, US, and South Korea with regard to transparency, privacy, and governance of AI-based IDS. This method will be used to find the differences and similarities with regard to the subject matter.

3. Technical Foundations of AI-Based Intrusion Detection Systems

3.1. IDS Fundamentals

Definition of IDS

An intrusion Detection System (IDS) is a security tool that can be implemented in either hardware or software. It monitors a network or system for malicious activity by analyzing network traffic, system activity, and carried data. When anomalous behavior is detected, the IDS generates alerts that provides metadata such as the detection time, source and destination IP addresses, and sometimes segments of packet payloads [2] – [17].

IDS vs. IPS

IDS should be distinguished from Intrusion Prevention Systems (IPS). An IDS does not stop a threat; it merely detects the suspicious activity and notifies administrators. On the other hand, an IPS not only detects the attack and alerts the administrator, but also intervenes by blocking the attack. IDS and IPS are usually used together, working as complementary parts of a layered security setup. Importantly, an IDS cannot use the same settings for all organizations, as it should be configured based on the specific requirements of that organization as uniform configuration cannot effectively handle different network setups.

Types of IDS: NIDS / HIDS / CIDS

IDS architectures can be classified based on their deployment location and what they monitor. According to [2], a Network-Based IDS (NIDS) monitors the entire network traffic and is usually deployed at strategic points such as the Demilitarized Zone DMZ, which allows it to have a broad view of network activity. It captures and analyzes inbound and outbound protocol packets from all devices in order to detect malicious traffic. Many organizations use NIDS together with firewalls and application-layer protection mechanisms to enhance web server security. A Host-Based IDS (HIDS) focuses on a specific host and monitors both inbound and outbound packets at the individual device level. It runs on servers, intermediate devices, and endpoints systems, and has access to both the internet and

the local network. HIDS monitors log files, running programs, and system files, and alerts the user when suspicious behavior occurs. It can also help prevent malware from spreading across the network. A Cloud-Based IDS (CIDS) is designed to monitor cloud environments and detect malicious behavior in cloud traffic. Together, NIDS and HIDS provide a layered defense strategy by covering both networkwide and host-level threats. However, traditional IDS approaches mainly rely on rule-based detection, which makes them less effective against new or evolving cyber threats [16]

Signature-based / anomaly-based / hybrid

Initially, IDS mainly used two categories of detection methods, which are Signature-Based and Anomaly-Based Detection. Signature-Based Detection uses a database of known attacks or a series of attacks. It matches the current network activity with the database. In case of a match, then the activity is considered a threat. Anomaly-Based Detection uses a different technique. First, normal behavior is identified, and then this information is used to identify unknown attacks, which are not normal. Third, there is the Hybrid Approach, which combines Signature-Based and Anomaly-Based Detection.

IDS tools examples

Several IDS tools are commonly used in practice.

SolarWinds Security Event Manager integrates log data from different parts of the IT infrastructure and analyzes them centrally, using both signature-based and anomaly-based detection methods.

McAfee LiveSafe combines a signature-based IPS to prevent known attacks with an anomaly-based IDS to detect and alert on unknown or unusual behavior.

Blumira monitors IT infrastructure for suspicious activities that could lead to data breaches and is able to respond to attacks in real time to stop malicious actors.

Evolution from Traditional IDS to AI-Based IDS

Traditional IDS limitations

Traditional IDS methods, including signature-based and rule-based approaches, are often inadequate against evolving attack techniques. Traditional intrusion detection systems cannot identify new threats if there is no existing known signature, while anomaly-based systems may fail to distinguish between malicious and legitimate anomalies, leading to a high number of false positives [16]

Traditional IDS rely on static rules, whereas modern attacks use irregular patterns [2]. [1] demonstrated that known Indicators of Compromise (IoCs) are commonly used in traditional threat detection methods, but conventional rule-based systems often fail to respond in real time. Additionally, [2] reported that the massive volume of data makes manual detection impossible, and traditional SOC operations require human analysts to monitor large amounts of network traffic continuously, manually detect anomalies, and generate reports, which is highly time-consuming and resource-intensive. Security managers usually try to prevent intrusions from external sources; however, web services in the internet DMZ (such as HTTPS, DNS, and FTP) must remain accessible, meaning firewalls cannot simply block them. Therefore, distinguishing legitimate access from malicious activity is essential, making intrusion detection a major research field in network security to protect internal networks from unusual access. One major trend that is reshaping the cybersecurity landscape is the development of AI-enhanced threat detection mechanisms.

Signature based limitations

Signature-based detection relies on a database of known attacks or predefined attack sequences and compares current network activity against this database. If a match is found, the activity is classified as a threat. However, traditional intrusion detection systems cannot identify new threats when no existing signature is available for them [16]. These systems also depend on static rules, while modern attacks often use irregular and evolving patterns. Furthermore, known Indicators of Compromise (IoCs) are widely used in traditional threat detection methods [1].

False Positives

Anomaly-based detection systems often struggle to distinguish between malicious and legitimate anomalies, which results in a high number of false positives [16]. To address this issue, AI can flag potentially malicious activities for further investigation instead of treating every deviation as a confirmed threat. In addition, AI can prioritize vulnerabilities based on their potential impact, allowing security personnel to focus on the most critical issues [1].

Concept Drift

[1] reported that Cyber threats are evolving in both complexity and speed, encompassing polymorphic malware and zero-day exploits. AI has the capability to continuously learn and adapt to new data and emerging threats, which enhances the robustness of security measures over time. AI-based intrusion detection systems leverage machine learning and deep learning to learn from these evolving attack behaviors; however, this intelligence also introduces unique ethical and governance challenges. Consequently, there is a growing need for autonomous and adaptive defense mechanisms.

Need for Adaptive Systems

Cybersecurity has changed a lot with the rise of AI. Instead of relying only on humans to watch over systems, AI now works alongside people, handling the heavy lifting like processing large amounts of data fast, while humans focus on understanding the context and making ethical decisions. AI can take over repetitive tasks, such as monitoring network traffic or analyzing alerts, which lets security teams work smarter and spend more time on complex problems [1]. In traditional Security Operations Centers (SOCs), analysts have to watch tons of network activity all the time, spot unusual behavior, and write reports, which is very time-consuming. AI-based intrusion detection systems (IDS) help reduce this workload by spotting threats quickly and letting humans focus on bigger decisions [2].

Machine learning (ML) and deep learning (DL) are at the core of modern IDS because they can handle huge datasets and detect complex patterns in real time [16]. These systems not only find anomalies but can also catch new types of attacks that older network- or host-based systems might miss. By monitoring user activity and network traffic constantly, AI-driven IDS can reduce the time between spotting a threat and responding to it, making the whole process faster and more efficient. AI also helps with automatic responses. When it detects a problem, AI systems can isolate infected devices, block suspicious traffic, and start fixing the issue almost immediately. Automated systems make it easier to contain threats and get systems back to normal quickly. Security Orchestration, Automation, and Response (SOAR) platforms help organize these processes so incidents are handled quickly and consistently. On the threat intelligence side, AI can make sense of data from different sources like threat feeds, social media, and even the dark web. With machine learning and Natural Language Processing (NLP), AI can scan reports, summarize updates, and find mentions of new malware or vulnerabilities. This helps security teams know what to focus on and respond faster. For example, Rapid7's Attack Intelligence Report [48] shows that 2023 had more zero-day vulnerabilities than ever before, which shows how fast threats are evolving. With the average cost of a data breach now at 4.88 million dollars [49], using AI for detection, response, and intelligence is becoming essential.

Why AI Became Necessary

There has been rapid growth in the integration of AI into the domain of cybersecurity. This has, in turn, changed the way organizations are dealing with the overall security of their digital assets. Thus, the integration of AI into the existing cybersecurity framework can be seen as a fundamental shift in the overall strategy. AI can process enormous amounts of data at extremely high speeds, thus helping identify anomalies and potential threats, which might be extremely difficult for humans to identify in real time. AI can also be used in conjunction with predictive analysis, where data can be used to predict potential future attacks [1].

Today, almost every industry requires data flow in order to function properly. This, in turn, leads us to an essential question regarding the overall safety of the data. With the rise of cyber-attacks and data breaches, failing to take proper care of the overall cybersecurity framework only leads to more chances of cyber-attacks. Although AI is being used at the highest levels in terms of productivity and innovation, it is also being used by cyber attackers to create sophisticated, self-encrypting malware and exploit code, which can be difficult to trace.

3.2. Machine Learning in IDS

Overview

Machine learning (ML) is a branch of artificial intelligence (AI) that allows systems to learn from data, recognize patterns, and make decisions with little to no human intervention. ML can be classified into two main categories: supervised learning and unsupervised learning. Various approaches: supervised, unsupervised, and semi-supervised, have been explored.

Supervised ML

In supervised learning, models are trained on data sets that are already labeled with the correct answers to enable the model to make predictions based on particular characteristics of the data. In the context of intrusion detection systems, this implies that the model will be trained on historical data that has been classified as normal or malicious and will use this to predict the outcomes based on particular characteristics of the data. In the context of cybersecurity, supervised learning is an important tool in the field because it can be used to solve classification problems. It can be used to classify whether incoming emails are legitimate or spam or whether particular network activities are good or bad. It can be used to classify data using decision trees, support vector machines (SVM), and logistic regression because it can recognize patterns in the data effectively [1].

Unsupervised ML

Unsupervised learning, on the other hand, does not rely on labeled data. Instead, it seeks to identify patterns and anomalies within the dataset without prior knowledge of what constitutes normal or malicious behavior [28].

Reinforcement Learning

According to [11], in contrast to traditional ML techniques, reinforcement learning (RL) aims to find an optimal learning policy strategy during the model-building phase.

Traditional ML relies upon an input and output pair, where an ML model receives a given event feature vector as input and outputs its estimated class value. RL-based approaches are significantly different. RL-based techniques are performed through an entity called the agent. RL-based techniques are not given the event class value. Instead, after an action is performed, the agent only receives the reward for its action and the subsequent environment state. Reinforcement learning has also been explored in IDS to enable adaptive and self-optimizing detection mechanisms. [10] propose a reinforcement learning-based intrusion detection model that maintains high detection accuracy over extended periods without frequent model updates by combining long-term learning with transfer learning, significantly reducing computational resources while outperforming traditional ML and deep learning schemes in F-measure.

Common Machine Learning Algorithms

Traditional ML methods including Artificial Neural Networks (ANN), Support Vector Machines (SVM), fuzzy approaches, swarm intelligence, and evolutionary computation techniques have been extensively employed for IDS. ANN: Inspired by the structure of the human brain with interconnected layers of neurons, can learn complex patterns from data and have proven effective for intrusion detection tasks [18]. SVM: One notable approach is the use of genetic algorithms combined with support vector machines (SVM) for intrusion detection. [9] used a genetic algorithm

to select key network features for an SVM classifier, improving anomaly-based IDS accuracy over traditional methods. Fuzzy Approaches: Leveraging fuzzy logic, these methods handle ambiguous or imprecise data, making them effective for analysing uncertain or noisy network traffic. Swarm Intelligence: Inspired by social insect behaviour, algorithms like Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO) optimize feature selection and rule generation for IDS [19].

Applications of Machine Learning in Intrusion Detection Systems

Using machine learning, IDS can learn patterns and create baselines of normal activity, known as trust models, to compare against current traffic. Numerous studies have explored this approach to develop IDS models that more efficiently detect security threats. For example, [8] proposed Passban, an anomaly-based IDS for IoT edge devices. By using edge computing and lightweight hardware accelerated logic gates, Passban identifies deviations from normal behavior directly on IoT gateways, showing how AI-based IDS can detect novel attacks in resource-limited environments beyond traditional SOC. Unlike signature-based IDS, it can catch unknown threats by spotting unusual patterns. In a more advanced application, Behavioral Analytics extends anomaly detection to user and entity behavior. Machine learning models create baselines of typical activity, such as login patterns, data access frequency, and activity contexts. By detecting deviations from these baselines, organizations can identify compromised accounts or insider threats that might evade conventional anomaly-based detection.

3.3. Deep Learning in IDS

Overview

Deep learning has received a lot of attention in recent years due to its ability to effectively identify intricate intrusion patterns within network traffic [53]. This is attributed to the fact that deep learning models consist of multiple layers, which learn increasingly abstract representations of data over time. This has, in turn, led to significant milestones in the field of computer vision, such as speech recognition [12]. This has led to the implementation of deep learning within intrusion detection systems, which can identify intricate patterns of intrusion within network traffic. As such, numerous cybersecurity solutions implemented over the years have incorporated deep learning, particularly within anomaly-based intrusion detection systems. For example, Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) have significantly contributed to the improvement of intrusion detection. Furthermore, deep learning models can identify spatiotemporal patterns within network traffic, which is often difficult for traditional models [26] – [27].

Most of the deep learning models used for intrusion detection purposes are trained on labeled network traffic datasets, such as CIC IDS, KDD Cup 99, NSL KDD, and UNSW NB15, which provide valuable insights for improving intrusion detection capabilities.

Core Architectures

CNN

Convolutional Neural Networks (CNNs) are commonly used for the spatial analysis of network traffic, while Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks are applied to learn temporal sequences in the data [13]. Previous studies have shown that, compared with other IDS classifiers, intrusion detection models based on CNNs often achieve higher detection rates and better precision [3].

RNN/LSTM

Recurrent Neural Networks (RNNs), particularly their advanced variant Long Short-Term Memory (LSTM) networks, are well suited for analyzing time-dependent data. Their ability to capture temporal relationships makes them useful for tracking network activity patterns and detecting potential intrusions through sequential behavior analysis. LSTM networks, in particular, are effective at handling sequential data, which allows them to identify temporal patterns and anomalies in network traffic that develop over time [15]. Some studies have also proposed RNN-based IDS models and compared them with other machine learning approaches such as ANN, Random Forest (RF), and SVM, showing that RNN-IDS models can achieve high classification accuracy [4]. In addition, hybrid approaches that combine CNNs and RNNs have shown promising results, as they leverage the strengths of both architectures to improve accuracy and detect more complex attack patterns [20].

DNN

While CNN and RNN architectures have shown strong performance, other deep learning approaches such as Deep Neural Networks (DNNs) have also been applied to IDS.

Several studies have proposed DNN-based IDS models, including [5], who validated their approach on benchmark datasets, and [6], who enhanced cloud-based anomaly detection using a hybrid optimization framework

Advantages

Deep learning models provide a vital advantage in intrusion detection systems in terms of their ability to learn features from data automatically. Models such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and even more recent models such as transformers have shown promising results in detecting unforeseen attack patterns in intrusion detection systems. Several researchers have also shown that these models can learn spatiotemporal features in network traffic flows, which is usually hard for traditional models to learn [26] – [27]. For instance, [5] designed a DNN-based IDS model for detecting unforeseen attacks in intrusion detection systems and evaluated their model using various benchmark datasets. The datasets used included NSL-KDD, UNSW-NB15, Kyoto, WSN-DS, and CICIDS2017. Using different datasets improved the accuracy of their model and made

it robust against various types of cyber-attacks. [6] designed a DNN model for detecting anomalies in cloud computing systems using a hybrid optimization technique involving an Improved Genetic Algorithm and Simulated Annealing.

It is, however, important to note that NLP is not widely used in traditional IDS, as the data used is mostly structured, such as network traffic data, which is not usually associated with NLP. Instead, other techniques such as decision trees, SVM, CNN, RNN, etc., are used for anomaly detection and classification of attacks based on numerical features of the data. However, there is some research on the use of NLP for IDS, which can be seen as a hybrid approach of combining NLP with other traditional IDS techniques for better understanding of network activities, rather than solely relying on numerical features of data for analysis.

Limitations and Deployment Challenges

Although these methods are accurate there are a number of challenges that need to be addresses.

Computational cost: These challenges include the computational cost of training these models using gradient descent-based algorithms such as Adam or RMSprop. In this respect, data augmentation, dropout, and batch normalization are often employed for the reduction of overfitting. However, the training of deep learning models for use in IDS systems is often associated with high computational costs that can only be addressed using high-performance hardware such as GPUs or TPUs [13]. Deep neural networks (DNNs), although powerful, face additional challenges as computational complexity grows with the number of hidden layers and neurons, which can make real-time detection less efficient. Recurrent architectures like LSTMs and RNNs are particularly useful for capturing long-term dependencies in network traffic but come with long training times due to their intricate structures. RNNs, in particular, can be computationally intensive and are prone to vanishing or exploding gradients, which can hinder performance on longer sequences [14].

Need for large labeled datasets: Supervised models achieve high accuracy but require large labeled datasets, whereas unsupervised methods can work with unlabeled data, often at the cost of lower accuracy. Semi-supervised techniques reduce the need for labeling but may perform inconsistently on smaller datasets. Despite their strengths, these approaches are limited because they depend on carefully labeled training data, which restricts their ability to generalize across different traffic types and attack scenarios and makes them inefficient due to the manual effort needed to label network flows [44]

Relying on only a few datasets may fail to capture all real-world intrusion scenarios, so training ML models on diverse datasets representing various types of network traffic and attack patterns allows IDS to generalize better and detect a broader range of threats. Using diverse datasets exposes models to multiple attack scenarios, though challenges such as data imbalance and domain adaptation must be addressed to fully benefit from this approach. For example, an XAI model demonstrated high predictive performance, achieving an accuracy of 87%, but it required the UNSW-NB15 dataset, which contains over 2.5 million records and nine different attack types [39].

Data imbalance: In AI-based IDS, data imbalance is one of the key issues that can significantly impact the performance of IDS models. Normally, the majority of data instances tend to be "normal" traffic, while anomalous attack data instances tend to be less frequent. This can cause the IDS model to be biased towards the majority-class instances during training, ignoring the minority-class instances. To deal with this problem, data imbalance handling techniques such as oversampling, undersampling, and Generative Adversarial Networks (GANs) have been proposed. For example, oversampling-based techniques, such as the Synthetic Minority Over-sampling Technique (SMOTE) [46], create new data instances to increase the size of the minority-class instances. Similarly, GANs have also been found to be promising for handling data imbalance, although training such models is known to be challenging, expensive, and requires careful hyperparameter adjustment. These techniques should be supported by large datasets [53].

Real-world scalability issues: While many AI-based IDS solutions perform well in labs or controlled environments, they often struggle when deployed in large, diverse, and constantly changing production networks. Combining new technologies, like edge computing, with existing infrastructure can create extra complexities and requires careful planning. Traditional setups may not handle increasing network traffic well, and making sure new AI tools integrate smoothly with current systems is another challenge. On top of that, staying effective against constantly evolving cyber threats means systems need regular updates, including threat intelligence feeds and periodic retraining of machine-learning models [28]. Many models also aren't fully ready for real-world conditions because they're usually tested on benchmark datasets, like KDD Cup 99, which don't reflect the full complexity and dynamics of live network traffic [2].

4. Ethical Challenges in AI-Based IDS

Despite impressive performance, ML and DL-based NIDS have shown over 95% detection accuracy in controlled tests [51], their real-world deployment faces important limitations. Most current models assume that training and testing data follow a static distribution, but network environments are constantly evolving. This concept drift, where the statistical behavior of both benign and malicious traffic changes over time, can cause models trained on original

datasets to misclassify or miss emerging threats. Beyond technical performance, these challenges raise important ethical questions: How transparent are the models' decisions? Could biases in the data lead to unfair or disproportionate detection? And how is sensitive network data being protected during training and operation? Exploring these issues is crucial for designing AI-based IDS that are not only accurate but also responsible and trustworthy

4.1. Transparency and Explainability

Problem

Machine learning-based intrusion detection systems (IDS) are usually characterized as “black boxes systems meaning that their decision-making process is opaque and difficult to understand. This lack of transparency not only limits trustworthiness but also raises ethical concerns around accountability and fairness. While state-of-the-art machine learning models are known to perform exceptionally well in various tasks, their level of complexity makes them less explainable. Consequently, a security analyst may find it difficult to validate or trust. Deep learning models, like DNN and Ensemble, work on the principle of learning complex patterns from large amounts of data. However, the inherent opacity of deep learning-based IDS systems poses a major ethical and operational challenges for security analysts. The black-box nature of deep learning models makes it hard for security analysts to assess the correctness of the detection alerts, since they do not have sufficient information on the features and factors affecting the model's decisions [22]. Evaluating the validity of IDS alerts and taking appropriate remedial actions requires clear insights into the rationale behind the system's decisions, yet black-box models rarely provide such transparency [39]. The complex architectures of deep learning models have added to the problem in terms of interpretability, with analysts struggling to understand how the system arrives at its conclusions [21]. The major issues in this regard have been due to the lack of interpretability in the models, which has limited analysts' ability to understand, trust, and act upon decisions made by the system and also raise ethical concerns, which include undetected errors, inability to audit decisions, and lack of accountability. In case any untoward or incorrect results are obtained, it has been found to be complex in terms of diagnosing the problem, which has led to sub-optimal performance or incorrect results going unnoticed. The problem has been compounded in industries like finance, health care, and government, which require transparency and accountability in decision-making processes. In such industries, black-box models have been found to be inadequate in meeting regulatory requirements, owing to the inability to audit explanations for predictions or actions made by the system [23].

Solutions:

To address the transparency issues of black-box models in IDS, a new technology called Explainable Artificial Intelligence (XAI) has been proposed. XAI is a field of research in artificial intelligence aimed at increasing the interpretability and trustworthiness of complex models, making them more comprehensible to humans [39] - [24]. The inclusion of XAI in forensic cybersecurity helps security professionals gain more insights from the models by providing clear explanations of model decisions, increasing their trust, enhancing technical validation, and strengthening ethical accountability, while also optimizing the performance of IDS. Recently, the field of XAI has gained more research, and various XAI technologies have been proposed. Among them, Local Interpretable Model-Agnostic Explanations (LIME) and Shapley Additive Explanations (SHAP) suggested by [37] have gained more recognition for their global and local interpretability capabilities.

LIME generates explanations for the black-box model by perturbing the instance and creating a weighted model of the nearest neighbors to the instance, providing a local approximation of the black-box model for a specific prediction. SHAP, on the other hand, assigns importance values to each feature using Shapley values from game theory, indicating positive or negative contributions while satisfying fairness properties such as Efficiency, Symmetry, Dummy, and Additivity. Incorporating XAI techniques into IDS pipelines can also help reduce false positives, enabling analysts to better validate flagged threats; for example, a 15% accuracy improvement has been reported when using LIME with ensemble models [42].

Weaknesses

Despite these advances, XAI methods face several limitations. LIME and SHAP explanations can be unstable under small input changes and may provide misleading attributions for complex temporal features [38]. Both techniques require significant computational resources, limiting their use in real-time IDS [39]. Moreover, the absence of standard evaluation criteria and measures also makes it hard to evaluate the reliability and reproducibility of the explanations [38]. XAI may also show correlations rather than cause-effect relations, which may lead to wrong interpretations in forensics. In the last few years, research on XAI has been increasing, and as a result, several techniques have been developed for XAI. Among all, Local Interpretable Model Agnostic Explanations (LIME) and Shapley Additive Explanations (SHAP) are notable for providing both global and local interpretability of the black box models. LIME achieves this by modifying the input and creating weighted interpretable models based on proximity to the instance, providing a local approximation of the black box for each prediction [37].

SHAP, despite its high theoretical accuracy, also faces major computational challenges. For instance, the approach demands the evaluation of the model several times to obtain the values for feature attribution. However, this results in high computation time, especially for the model-agnostic approach, KernelSHAP, which is known to have an exponential increase in computation with the number of features [38][39]. Therefore, the approach is less applicable for real-time intrusion detection systems, making the approach more appropriate for post-intrusion analysis. Moreover, the approach is mathematically complex, making it hard for non-experts to understand. Additionally, the approach is limited in its applicability to high-dimensional data, especially for deep learning models, as obtaining values or explanations can be computationally expensive, making the approach challenging [39].

LIME provides faster and more intuitive explanations compared to SHAP, making it useful for quick diagnostics and exploratory analysis. However, its results are highly sensitive to the perturbation technique and sample generation process, meaning repeated analyses of the same instance can yield different explanations. LIME also lacks formal guarantees such as completeness or consistency, which reduces its reliability in legal or forensic reporting. Moreover, its reliance on local linear approximations may misrepresent complex model behavior, particularly in deep learning contexts. While these characteristics make LIME convenient for initial investigations, it remains unsuitable for applications that demand high stability, fidelity, and legal defensibility.

Unsolved Problems / Research Gaps

Many studies have explored explainable AI (XAI) in intrusion detection systems (IDS), focusing on how these methods can improve interpretability and effectiveness. However, a major challenge remains: the lack of standardized metrics and benchmarks to evaluate explanation quality specifically for IDS tasks. Although numerous works propose XAI techniques, there is currently no universally accepted framework for assessing robustness, stability, or domain-specific utility in security analysis. Existing surveys highlight this absence of agreed-upon evaluation metrics as a significant gap in explainable intrusion detection research [54].

The main challenge in implementing XAI for network intrusion detection lies in validating these tools, testing their quality, and evaluating relevant security metrics. Without universally accepted metrics, it becomes difficult to compare and benchmark different XAI methods. Reliable metrics are crucial, particularly in high-stakes domains, to prevent misleading explanations and ensure that outputs can be trusted for securitycritical decision-making [55].

Another important research gap exists at the intersection of XAI performance and deployment feasibility. Optimizing XAI for high-frequency and resource-constrained network environments is critical. Popular post-hoc techniques like SHAP and LIME, while effective for explanation, come with computational costs that hinder real-time applicability in high-traffic networks. For example, SHAP provides comprehensive feature attributions but is often too slow for rapid deployment, especially in high-speed networks where latency must be minimized [39]. LIME is somewhat more efficient but still presents challenges in such environments.

Recent systematic analyses [56] highlight that the impact of XAI on detection performance and computational efficiency in resource-constrained IoT networks is still not fully understood. These studies show a significant trade-off between achieving high explainability and maintaining computational efficiency, which can make real-world deployment infeasible. Furthermore, the lack of standardized reporting of computational metrics remains a major obstacle to practical progress. Addressing these gaps, through standardized evaluation, efficiency optimization, and careful deployment design will be essential to ensure that XAI methods enhance both the technical and ethical performance of IDS in real-world settings.

4.2. Bias and Fairness

Problem

One of the major ethical and technical challenges in AI-based IDS is bias, often resulting in high false-positive or false-negative rates. This problem largely stems from imbalanced datasets, where attack instances are far fewer than benign traffic, causing models to favor the majority class. Despite advances in machine learning, many systems remain biased because most available data represent normal behavior, and attack data are scarce. Class imbalance directly contributes to incorrect detection, while the laborintensive process of labeling network traffic, especially evolving or rare attacks, makes it difficult to prepare balanced datasets for effective adaptation [33] – [34].

False alarms are another consequence of this bias. Anomaly-based AI IDS models often learn only normal behavior, so any harmless change in normal traffic may be misclassified as malicious. This not only wastes IDS resources but also reduces overall system effectiveness. Traditional misuse detection relies on predefined signatures, which work well for known attacks but fail for unknown ones, often producing high false-positive rates similar to anomaly-based approaches [2]. Consequently, human analysts are still required to verify AI alerts, limiting operational efficiency and raising ethical concerns regarding trust and accountability in automated intrusion detection.

Solutions

To mitigate bias and class imbalance, several strategies have been proposed. Conventional techniques include undersampling and oversampling [36] – [45]. Random Undersampling (RUS) is a widely used technique to address class imbalance in machine learning datasets. For instance, it balances datasets by randomly removing samples

from the majority class, which can reduce bias toward normal traffic [47]. Oversampling methods, such as SMOTE (Synthetic Minority Over-sampling Technique), generate synthetic samples for minority classes by interpolating existing instances, improving representation in training data [46].

More sophisticated approaches have also emerged. PROVSYN (2025) is an automated framework that synthesizes provenance graphs using graph neural networks, large language models, and rule-based refinement. By generating realistic attack data, it balances normal and malicious samples, ensuring underrepresented attack types are adequately represented. This augmentation improves detection accuracy, reduces false positives, and enriches underrepresented labels, particularly in smaller datasets like Nodlink [43].

Self-supervised transformer-based contrastive learning techniques have also shown promise [44]. These models learn useful representations from both common and underrepresented classes, reducing reliance on imbalanced labeled data. Using packet-sequence and flow-statistics datasets, transformer encoders trained with contrastive learning can capture minority or rare patterns, improving generalization across diverse and imbalanced distributions. This approach leverages a BERT-based architecture and a contrastive learning process that maximizes similarity between related data points while minimizing it between dissimilar ones, ensuring that even rare attack patterns are effectively represented.

Generative active adaptation methods, such as NetGuard [50], further address bias by combining density-aware dataset selection with conditional data augmentation. These frameworks help NIDS adapt to concept drift and class imbalance, generating more diverse data and enhancing model robustness for rare attacks like web attacks and botnets. NetGuard identifies dataset priors and synthesizes new data conditioned on them, allowing models to generalize better when training and testing distributions differ over time. The framework has been tested on both simulated IDS data and real-world ISP datasets, demonstrating significant improvements in intrusion detection performance in dynamic environments while reducing labeling effort and adaptation costs. By prioritizing the most informative samples, NetGuard also alleviates the challenges of uncertainty-based active learning, helping the system respond faster to evolving threats.

Weaknesses

Although the solutions provided can help mitigate Bias, limitations remain. Undersampling methods like RUS risk data loss in small datasets, potentially discarding important information and causing overfitting. Random selection may also remove critical patterns, leading to inconsistent results. Oversampling techniques, such as SMOTE, may introduce noise and fail to capture complex data structures, sometimes amplifying borderline cases or class overlap. While PROVSYN overcomes some of these issues, it still approximates temporal modeling rather than explicitly representing timestamps, generates smaller-than-realistic graphs, and relies on sequential text attribute generation that limits efficiency in larger graphs. Its applicability to other domains or downstream algorithms also remains underexplored. Generalization to real-world environments remains a challenge. Most solutions, including PROVSYN, SSL-based models, and NetGuard, are evaluated on specific benchmark datasets (e.g., CIDDS, NSL-KDD, Nodlink) that are either outdated or artificially balanced. Consequently, these models often fail to generalize to unseen networks or novel attack types. Additionally, bias mitigation models often lack transparency and explainability. SSL and transformer-based IDS models, as well as PROVSYN, remain black boxes. They improve balance and detection accuracy but provide little insight into how minority classes are represented or how bias is reduced. This opacity limits trust and makes it difficult to identify hidden biases learned during pretext tasks [44] – [53].

Unsolved Problems / Research Gaps:

Future work in this direction would involve the integration of Explainable AI (XAI) with imbalance-aware and generative techniques to provide informative and meaningful knowledge about the model's behavior, enabling security experts to know the features used by the SSL model to represent attacks. There is a lack of a framework to interpret and visualize the representations learned by the SSL model in the context of IDS. Moreover, the authenticity of the synthetic data used in the provenance needs to be validated, and metrics should be developed to assess the realism and diversity of the generated data [50]. Moreover, there is a need to develop self-evolving models that are capable of learning and adapting to concept drift and rebalancing the minority samples in near-real-time, which is still an open problem in the field [52].

All these issues are critical in the context of ensuring that the mitigation of biases in IDS systems is not only effective in improving the system's performance but also ensures the ethical and moral integrity of the system.

4.3. Privacy and Data Protection

Problem

In the case of Intrusion Detection Systems (IDS) using AI, there is a need to collect large amounts of data in the form of logs and telemetry data, including packet captures, flow records, and host logs, which may contain personal data. The use of IDS involves the analysis of packet headers and sometimes payload data, which is a form of surveillance and breach of user confidentiality. This involves the continuous monitoring of the network, the system, and the user. Since the Intrusion Detection System learns patterns, large amounts of data collection are necessary. For the purpose of experiments, packet captures in the form of pcap, NetFlow, and benchmark data are used, and the model cannot work without data.

However, there are some challenges that are associated with the use of data, as the amount of data collected can

rise to terabytes per day, thus posing a problem for storage. Network-based IDS use the flow of data passing through the network device to detect intrusions, while host-based IDS use the processes and files accessed on a single device to reveal patterns of user behavior. For better anomaly detection, the IDS have to use data from the kernel level and the network level, sometimes combining the data from the kernel level and the network level for better performance [57]. Deep learning-based methods, such as the Deep defense model, have been seen to produce better results than traditional machine learning methods, especially when the data set is large [58]. These methods use the data collected from the network, such as the data from the packets, to produce better performance, thus the suitability of the methods for use with the extensive data passing through the network

Altogether, AI-based IDS rely on extensive data collection and deep system visibility, making privacy monitoring and processing challenging. Inspecting network traffic and file activity in detail can reveal patterns of user behavior, increasing privacy risks.

Solutions

One promising solution is Federated Learning (FL). Centralizing network data for AI model training can violate privacy regulations such as GDPR [59]. FL has emerged as a privacy-preserving approach that allows distributed model training across multiple IoT devices without centralizing data [29]. Devices train local models independently and share only model updates with a central server, minimizing privacy risks while enabling collaborative model learning. The central server then aggregates updates and redistributes them to participants. Hybrid cryptographic approaches further enhance privacy: for instance, homomorphic encryption (HE) can secure data offloaded to a central server, allowing the server to learn from encrypted data without accessing its content. This maintains confidentiality and improves efficiency in FL-based IDS, such as in IoV networks [30].

Another approach is Differential Privacy (DP). In federated learning settings, sharing model gradients or updates may expose sensitive data through gradient leakage or model inversion attacks [31]. DP is a mathematical privacy-preserving technique that mitigates this by adding carefully calibrated noise to data or model updates before sharing, reducing the risk of reconstructing original data. It provides mathematical privacy guarantees, making it difficult for adversaries to infer individual network traffic records or user behavior while enabling collaborative learning across organizations without exposing raw data.

Weaknesses

Although such advancements have been made, it is not entirely flawless, and this has led to ethical issues. When it comes to AI-based IDS, where gradients are computed based on sensitive network traffic logs, there is still room for privacy leakage, even if it is decentralized. Distributed training and collaborative learning are widely adopted in privacy-sensitive scenarios, where raw data never leaves local devices. In such systems, only gradients are exchanged, and it is commonly assumed that this protects participants' training data. However, research challenges this assumption, demonstrating that gradients can leak sensitive information, which raises serious ethical questions about user consent and data confidentiality. The authors explicitly question whether gradients truly protect privacy and show that it is possible to completely reconstruct the original training data from shared gradients. They propose a novel attack called Deep Leakage from Gradients (DLG), an optimization-based method that reconstructs training inputs and labels using only the gradients. DLG works by optimizing dummy inputs and labels to minimize the distance between the dummy gradients and the original gradients. Remarkably, the authors can achieve pixel-perfect images and token-perfect texts without any generative model or prior knowledge of the data. Experimental results show that the reconstructed samples are almost identical to the original training data, highlighting a critical ethical challenge: even sharing seemingly anonymized gradients can compromise user privacy [60].

Computation and performance limitations also pose ethical implications [61]. Fully Homomorphic Encryption (FHE), such as Gentry's ideal lattice-based FHE [Gentry 2009], supports any number of computations over the data. However, the high computational overhead and mathematical complexity of FHE make its practical implementation challenging. This leads to ethical issues in terms of fairness, as some organizations cannot afford the required infrastructure for privacy. The bootstrapping operation is very expensive for FHE, though some leveled FHE provides a solution for improving performance. Researchers continue to explore somewhat homomorphic encryption (SWHE) schemes that balance security and practicality, but applying them in real-time or large-scale IDS scenarios remains an open problem.

Additionally, privacy-preserving methods like Differential Privacy introduce a privacy–utility trade-off, which carries both technical and ethical implications [32]. While adding noise to gradients or data helps protect sensitive information, overly conservative noise can significantly degrade model performance, potentially affecting the system's reliability. Factors like batch size and noise scale (σ) significantly affect the accuracy of the model. If not tuned properly, the accuracy of the model will be completely destroyed. From an ethical perspective, the designers of privacy-preserving IDS must balance the moral responsibility to protect sensitive user data with the need to maintain effective system performance, highlighting an ongoing tension between security, usability, and ethical accountability.

Unsolved Problems / Research Gaps

Practical, scalable integration of HE and FL for resource-limited IDS remains a challenge [62]. Existing solutions often suffer from high computational and communication overheads. Even optimized HE schemes require further validation to be feasible in real-world scenarios, such as real-time recommender systems. Communication costs remain high for large-scale applications.

Another gap is robust privacy under non-IID data distributions [63]. the training data across clients is often not IID, resulting in the divergence of the model weights during the training of the FL models. Attackers can also extract sensitive information from the model parameters in such settings, thus the need for privacy-preserving mechanisms that are robust even in the presence of non-IID data distributions. Finally, the importance of governance and compliance mapping cannot be overemphasized. The data minimization, purpose limitation, privacy by design, and risk management principles of the GDPR, CCPA, and NIST Privacy Framework are all relevant in the context of data privacy. The use of data sheets, model cards, and privacy budget ledgers provides a measure of accountability and transparency. While the use of FL is technically compliant with the regulations, the need for the use of DP and aggregation mechanisms to address the privacy risk further underscores the ethical responsibility to ensure enforceable, transparent, and robust protections in AI-based IDS deployments.

5. Comparative Analysis of EU, U.S., and South Korea

5.1. Transparency and Explainability in AI-Based Intrusion Detection Systems

The European Union, the United States, and South Korea approach transparency in AI-based intrusion detection systems (IDS) differently, reflecting their distinct governance priorities and organizational philosophies.

European Union: Explainability is considered a legal requirement. According to [64], transparency is considered a matter of fundamental rights under the General Data Protection Regulation (GDPR). There are strict documentation and risk assessment requirements for high-risk AI systems, such as those for IDS in CI [65]. Black box IDS models, which are unable to provide any kind of explanation, could potentially face issues in terms of compliance, particularly if decisions are automatically made and have implications for individual rights [66].

United States: Explainability is approached as a governance and risk management practice [67]. Unlike the EU, the emphasis is not on the management of operational risks but rather the enforcement of the rights-based approach. Therefore, even though NIST's privacy guidelines emphasize the importance of transparency and accountability, they do not emphasize the importance of explainability. Thus, SHAP and LIME are best practices, but it is not necessarily driven by legal requirements.

South Korea: The approach finds a delicate balance between data protection and national cybersecurity concerns. The Personal Information Protection Act (PIPA) has robust data protection principles, similar to the GDPR [68]. The requirements for AI explainability are not as strong, however [69]. The national cybersecurity strategy, which involves various actors such as the Korea Internet & Security Agency (KISA), prioritizes the technological security of the country and the swift deployment of AI [70]. Transparency is encouraged, but functionality is prioritized. Such differences give rise to asymmetric compliance pressures. For example, a black box IDS deployed in a U.S. enterprise may not comply with EU regulations for high-risk AI due to detailed auditing requirements, whereas explanation requirements that comply with EU regulations may not satisfy latency requirements in South Korea's cybersecurity landscape.

5.2. Privacy and Data Governance: Centralization vs. Federated Approaches

AI-based intrusion detection systems rely heavily on access to large volumes of network and user data. How this data is collected, stored, and processed varies across regions, reflecting different regulatory philosophies and cybersecurity priorities

European Union: GDPR principles, such as data minimization, limitation of purposes, and privacy by design, impose strong constraints on centralized data aggregation. Training AI models on raw network logs requires Data Protection Impact Assessments (DPIAs). Federated learning (FL) architectures are better aligned with EU regulations, as they keep data local and limit raw data sharing [64].

United States: Privacy rules are sector-specific rather than unified. Some industries, like healthcare and finance, have strict data-handling regulations, but cybersecurity applications allow more flexibility. Centralized data collection for IDS training is generally permitted if safeguards are in place. Federated learning is often adopted for efficiency or innovation rather than regulatory necessity [67].

South Korea: PIPA enforces strict consent and data protection rules. However, the country's centralized digital infrastructure and government-driven cybersecurity efforts allow large-scale data processing for national defense if safeguards exist. Federated learning may be used to enhance trust and security, but decentralization requirements are not as strong as in the EU [70].

Across all regions, federated learning alone does not eliminate privacy risks. Techniques like gradient leakage demonstrate that decentralizing data is not enough. Compliance depends on additional protections such as differential privacy or secure aggregation.

5.3. Bias, Fairness, and Operational Trade-Offs

Ethical considerations in AI-based IDS vary significantly across regions, reflecting different regulatory frameworks, societal values, and operational priorities.

European Union: Fairness and mitigation of risk are integral to the compliance measures for high-risk AI. PROVSYN or self-supervised learning can be employed for this [65].

United States: Fairness in AI policy is gaining traction, but cybersecurity focuses more on the efficiency of operations. False positive reduction is viewed as an improvement to the effectiveness of the SOC rather than a user right [67] – [71].

South Korea: AI guidelines emphasize trust and social responsibility, yet cybersecurity resilience is the top priority. Bias reduction in IDS is generally evaluated based on its impact on detection accuracy and national cyber defense [70].

In practice, the EU may accept slightly lower detection accuracy to better protect user rights, while in the U.S. and South Korea, operational efficiency and detection effectiveness often take precedence over privacy or explainability concerns.

6. Recommendations

On the basis of the analysis provided within this study, some recommendations could be proposed to enhance the ethical design and deployment of AI-based intrusion detection systems. Firstly, more emphasis should be given to the inclusion of explainable artificial intelligence (XAI) techniques within the IDS models. Increasing the transparency of the model could help security analysts to understand the decision process and detect possible errors within the detection process. Further research should focus on the development of efficient XAI techniques with the ability to provide explanations without affecting the overall computational cost. Second, the problem of dataset imbalance and bias should continue to receive significant attention in the field of IDS. Data generation techniques, data augmentation methods, and self-supervised learning methods can all contribute to the mitigation of the problem of dataset imbalance. On the other hand, evaluation standards need to be established to test the effectiveness of fairness mitigation techniques in intrusion detection systems.

Third, privacy-preserving techniques need to be incorporated in the training of AI models to handle network data. Techniques like federated learning, differential privacy, and secure aggregation can minimize the need for data collection in the training of AI models. These techniques need to be fine-tuned to minimize the computational overhead in the training of AI models. Fourth, organizations need to develop a set of governance frameworks that support the alignment of the deployment of AI-based IDS with ethical and regulatory requirements. Documentation methods like model cards, data sheets, and risk reports can be useful. Moreover, the implementation of international privacy regulations should be integrated into the design phase of AI systems rather than the post-deployment phase. Lastly, the collaboration of the academic community with the industry and the government is a key factor to develop standardized benchmarks for the evaluation of AI-based IDS to ensure that the AI systems deployed for intrusion detection achieve high performance with ethical considerations.

7. Conclusion

Artificial intelligence has helped intrusion detection systems to make them quicker and better at recognizing complex cyber intrusions. Machine learning and deep learning techniques have helped intrusion detection systems to better analyze the data from the network and recognize patterns. With the increasing sophistication of cyberattacks, AI-based intrusion detection systems have started to play an important part in the cybersecurity of organizations. However, the use of AI in intrusion detection systems has raised various ethical issues. Some of the issues include limited transparency, bias in the data used for training the AI model, and the use of collected data for breaching privacy. These issues have raised questions of trust in the use of AI in intrusion detection systems. In addition, different parts of the globe may have different strategies for the governance of AI. For example, the European Union may place more importance on the aspect of privacy and transparency, whereas the USA, as well as South Korea, may place more importance on the aspect of operational efficiency, as well as cybersecurity resilience. In the future, it is important to work towards enhancing the aspect of transparency, fairness, and privacy, as well as detection. This will require the efforts of different parties, including researchers, policymakers, as well as the industry, to develop effective intrusion detection systems based on artificial intelligence.

REFERENCES

- [1] Khan, M. I., Arif, A., & Khan, A. R. A. (2024). "The most recent advances and uses of AI in cybersecurity". *BULLET: Jurnal Multidisiplin Ilmu*, 3(4), 566–578. <https://journal.mediapublikasi.id/index.php/bullet/article/view/4540>.
- [2] Kim, A., Park, M., & Lee, D. H. (2020). AI-IDS: Application of deep learning to real-time web intrusion detection. *IEEE Access*, 8, 70245–70261 <https://doi.org/10.1109/ACCESS.2020.2986882>
- [3] Liu, Y., Liu, S., & Zhao, X. (2017). Intrusion detection algorithm based on convolutional neural network. *Beijing Ligong Daxue Xuebao/Transactions of Beijing Institute of Technology*, 37(12), 1271–1275. <https://doi.org/10.15918/j.tbit1001-0645.2017.12.011>
- [4] Yin, C., Zhu, Y., Fei, J., & He, X. (2017). A deep learning approach for intrusion detection using recurrent neural

- networks. *IEEE Access*, 5, 21954–21961. <https://doi.org/10.1109/ACCESS.2017.2762418>
- [5] Vinayakumar, R., Alazab, M., Soman, K. P., Poornachandran, P., Al-Nemrat, A., & Venkatraman, S. (2019). Deep learning approach for intelligent intrusion detection system. *IEEE Access*, 7, 41525–41550. <https://doi.org/10.1109/ACCESS.2019.2895334>
 - [6] Chiba, Z., Abghour, N., Moussaid, K., El Omri, A., & Rida, M. (2019). Intelligent approach to build a deep neural network-based IDS for cloud environment using combination of machine learning algorithms. *Computers & Security*, 86, 291–317. <https://doi.org/10.1016/j.cose.2019.06.009>
 - [7] Muneer, S., Farooq, U., Athar, A., Raza, M. A., Ghazal, T. M., & Sakib, S. (2024). A critical review of artificial intelligence based approaches in intrusion detection: A comprehensive analysis. *Journal of Cybersecurity Research*, 15 April. <https://doi.org/10.1155/2024/3909173>
 - [8] Eskandari, M., Janjua, Z. H., Vecchio, M., & Antonelli, F. (2020). Passban IDS: An intelligent anomaly-based intrusion detection system for IoT edge devices. *IEEE Internet of Things Journal*, 7(8), 6882–6897. <https://doi.org/10.1109/JIOT.2020.2970501>
 - [9] Gharaee, H., & Hosseinvand, H. (2016). A new feature selection IDS based on genetic algorithm and SVM. In 2016 8th International Symposium on Telecommunications (IST) (pp. 139–144). IEEE. <https://doi.org/10.1109/ISTEL.2016.7881798>
 - [10] Santos, R. R. d., Viegas, E. K., Santin, A. O., & Cogo, V. V. (2023). Reinforcement learning for intrusion detection: More model longness and fewer updates. *IEEE Transactions on Network and Service Management*, 20(2), 2040–2055. <https://doi.org/10.1109/TNSM.2022.3207094>
 - [11] Wassermann, S., Cuvelier, T., Mulinka, P., & Casas, P. (2021). Adaptive and reinforcement learning approaches for online network monitoring and analysis. *IEEE Transactions on Network and Service Management*, 18(2), 1832–1849. <https://doi.org/10.1109/TNSM.2020.3037486>
 - [12] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521, 436–444. <https://doi.org/10.1038/nature14539>
 - [13] Tomide, A., & Adelusi, J. (2025). Deep learning for intrusion detection: Challenges and opportunities. *Journal of AI Security Research*, 10(2), 45–62.
 - [14] Chinnasamy, R., Subramanian, M., Easwaramoorthy, S. V., & Cho, J. (2025). Deep learning-driven methods for network-based intrusion detection systems: A systematic review. *ICT Express*, 11(1), 181–215. <https://doi.org/10.1016/j.ict.2025.01.005>
 - [15] Fu, X., Zhou, N., Jiao, L., et al. (2021). The robust deep learning-based schemes for intrusion detection in Internet of Things environments. *Annals of Telecommunications*, 76, 273–285. <https://doi.org/10.1007/s12243-021-00854-y>
 - [16] Mohale, V. Z., & Obagbuwa, I. C. (2025). Evaluating machine learning-based intrusion detection systems with explainable AI: Enhancing transparency and interpretability. *Frontiers in Computer Science*, 7, 1520741. <https://doi.org/10.3389/fcomp.2025.1520741>
 - [17] Asharf, J., Moustafa, N., Khurshid, H., Debie, E., Haider, W., & Wahab, A. (2020). A review of intrusion detection systems using machine and deep learning in Internet of Things: Challenges, solutions and future directions. *Electronics*, 9, 1177. <https://doi.org/10.3390/electronics9071177>
 - [18] Khatkar, M., Kumar, K., & Kumar, B. (2023). Design and analysis of intrusion detection system based on ensemble deep neural network and XAI. *International Review of Modelling and Simulation*, 16, 129. <https://doi.org/10.15866/iremos.v16i3.23437>
 - [19] Balyan, A. K., Ahuja, S., Lilhore, U. K., Sharma, S. K., Manoharan, P., Algarni, A. D., et al. (2022). A hybrid intrusion detection model using EGAPSO and improved random forest method. *Sensors*, 22, 5986. <https://doi.org/10.3390/s22165986>
 - [20] Mynuddin, M., Khan, S. U., Chowdhury, Z., Islam, F., Islam, M., Hossain, M., et al. (2024). “Automatic network intrusion detection system using machine learning and deep learning”. [Preprint]. <https://doi.org/10.36227/techrxiv.170792293.35058961/v1>
 - [21] Barnard, P., DaSilva, L. A., & Marchetti, N. (2024). Don't just explain, enhance! Using explainable artificial intelligence (XAI) to automatically improve network intrusion detection [Preprint]. *arXiv*. <https://arxiv.org/abs/2407.17373>
 - [22] Neupane, S., Ables, J., Anderson, W., Mittal, S., Rahimi, S., Banicescu, I., et al. (2022). Explainable intrusion detection systems (X-IDS): A survey of current methods, challenges, and opportunities. *IEEE Access*, 10, 112392–112415. <https://doi.org/10.1109/ACCESS.2022.3216617>
 - [23] Sevri, M., & Karacan, H. (2023). Explainable artificial intelligence (XAI) for deep learning-based intrusion detection systems. In 4th International Conference on Artificial Intelligence and Applied Mathematics in Engineering (pp. 39–55) https://doi.org/10.1007/978-3-031-31956-3_4
 - [24] Wang, L., Zhang, Y., & Li, H. (2024). Explainable AI-based innovative hybrid ensemble model for intrusion detection systems. *Journal of Cloud Computing: Advances, Systems and Applications*, 13, 71. <https://doi.org/10.1186/s13677-024-00712-x>
 - [25] Zhang, Y., Muniyandi, R. C., & Qamar, F. (2025). A review of deep learning applications in intrusion detection systems: Overcoming challenges in spatiotemporal feature extraction and data imbalance. *Applied Sciences*, 15(3), 1552. <https://doi.org/10.3390/app15031552>
 - [26] Sharma, B., Sharma, L., Lal, C., & Roy, S. (2023). Anomaly based network intrusion detection for IoT attacks using deep learning technique. *Computers & Electrical Engineering*, 107, 108626.

- <https://doi.org/10.1016/j.compeleceng.2022.108626>
- [27] Shu, X., Zhang, L., Qi, G. J., Liu, W., & Tang, J. (2022). Spatiotemporal co-attention recurrent neural networks for human-skeleton motion prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44, 3300–3315. <https://doi.org/10.1109/TPAMI.2021.3099213>
- [28] Raja, M. S. (2025). The rise of AI-driven network intrusion detection systems: Innovations, challenges, and future directions. *International Journal of AI, Big Data, Computational and Management Studies*, 6(1), 1–9. <https://doi.org/10.63282/3050-9416.IJAIDCMS-V6I1P101>
- [29] Ansam Khraisat, Ammar Alazab, Sarabjot Singh, Tony Jan, and Alfredo Jr. Gomez. 2024. Survey on Federated Learning for Intrusion Detection System: Concept, Architectures, Aggregation Strategies, Challenges, and Future Directions. *ACM Comput. Surv.* 57, 1, Article 7 (January 2025), 38 pages. <https://doi.org/10.1145/3687124>
- [30] Bui Duc Manh, Chi-Hieu Nguyen, Dinh Thai Hoang, Diep N. Nguyen. 2024 Homomorphic Encryption-Enabled Federated Learning for Privacy-Preserving Intrusion Detection in Resource-Constrained IoT Networks <https://doi.org/10.48550/arXiv.2407.18503>
- [31] Manzoor, H. U., Shabbir, A., Chen, A., Flynn, D., & Zoha, A. (2024). A Survey of Security Strategies in Federated Learning: Defending Models, Data, and Privacy. *Future Internet*, 16(10), 374. <https://doi.org/10.3390/fi16100374>
- [32] Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* <https://doi.org/10.1145/2976749.2978318>
- [33] Mikołajczyk-Bareła, A., Ferlin, M., & Grochowski, M. (2023, August 23). Targeted data augmentation for bias mitigation. *Gdańsk University of Technology*. <https://doi.org/10.48550/arXiv.2308.11386>
- [34] Gurung, K., Ghimire, A., & Amsaad, F. (26 July 2025). Enhancing IoT intrusion detection systems through adversarial training. *Wright State University*. <https://doi.org/10.48550/arXiv.2507.19739>
- [35] Marek Pawlicki, Aleksandra Pawlicka, Rafał Kozik, Michał Choraś, A metasurvey of adversarial attacks against artificial intelligence algorithms, including diffusion models, *Neurocomputing*, Volume 653, 2025, 131231, ISSN 0925-2312, <https://doi.org/10.1016/j.neucom.2025.131231>.
- [36] Pastaltzidis, I., Dimitriou, N., Quezada-Tavárez, K., Aidinis, S., Marquenie, T., Gurzawska, A., & Tzovaras, D. (2022-06-20). Data augmentation for fairness-aware machine learning: Preventing algorithmic bias in law enforcement systems. *Centre for Research and Technology Hellas*. <https://doi.org/10.1145/3531146.3534644>
- [37] Gaspar, D., Silva, P., & Silva, C. (2024). Explainable AI for intrusion detection systems: LIME and SHAP applicability on multi-layer perceptron. *IEEE Access*, 1–1. <https://doi.org/10.1109/ACCESS.2024.3368377>
- [38] Hermosilla, P., Berrios, S., & Allende-Cid, H. (2025). Explainable AI for Forensic Analysis: A Comparative Study of SHAP and LIME in Intrusion Detection Models. *Applied Sciences*, 15(13), 7329. <https://doi.org/10.3390/app15137329>
- [39] Mohale, V. Z., & Obagbuwa, I. C. (2025). A systematic review on the integration of explainable artificial intelligence in intrusion detection systems to enhancing transparency and interpretability in cybersecurity. *Frontiers in artificial intelligence*, 8, 1526221. <https://doi.org/10.3389/frai.2025.1526221>
- [40] M. T. Ribeiro, S. Singh and C. Guestrin, "Why should I trust you?" Explaining the predictions of any classifier, *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016. <https://doi.org/10.48550/arXiv.1602.04938>
- [41] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions [Preprint]. *arXiv*. <https://doi.org/10.48550/arXiv.1705.07874>
- [42] Kotecha, K., Dhupar, V., Chandra, S., & Kumar, M. (2022). Explainable artificial intelligence for intrusion detection system. *Electronics*, 11, 3079. <https://doi.org/10.3390/electronics11193079>
- [43] Huang, Y., Ul Hassan, W., Guo, Y., Chen, X., & Li, D. (2025, June 6). PROVSYN: Synthesizing provenance graphs for data augmentation in intrusion detection systems. *arXiv preprint arXiv:2506.06226*. <https://arxiv.org/abs/2506.06226>
- [44] Koukoulis, I., Syrigos, I., & Korakis, T. (2025, May 12). Self-supervised transformerbased contrastive learning for intrusion detection systems. *Department of Electrical and Computer Engineering, University of Thessaly; Centre for Research and Technology Hellas (CERTH)*. <https://doi.org/10.48550/arXiv.2505.08816>
- [45] Mohammed, R., Rawashdeh, J., & Abdullah, M. (2020). Machine learning with oversampling and undersampling techniques: Overview study and experimental results. In *2020 International Conference on Information and Communication Technology for Intelligent Systems (ICICTIS)* (pp. 243–248). *IEEE*. <https://doi.org/10.1109/ICICTIS49469.2020.239556>
- [46] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321–357. <https://doi.org/10.1613/jair.953>
- [47] Hasanin, T., & Khoshgoftaar, T. (2018). The effects of random undersampling with simulated class imbalance for big data. In *2018 IEEE International Conference on Information Reuse and Integration (IRI)* (pp. 70–79). *IEEE*. <https://doi.org/10.1109/IRI.2018.00018>
- [48] Rapid7. (2024). 2024 attack intelligence report. <https://www.rapid7.com/research/report/2024-attack-intelligence-report/>

- [49] Cost of a Data Breach Report 2024. [Online]. Available: <https://www.ibm.com/reports/data-breach>
- [50] Gupta, R., Liu, S., Zhang, R., Hu, X., Wang, X., Benkraouda, H., Kommaraju, P., Cao, P., Feamster, N., & Nahrstedt, K. (n.d.). Generative active adaptation for drifting and imbalanced network intrusion detection. University of Illinois Urbana-Champaign and University of Chicago. <https://doi.org/10.48550/arXiv.2503.03022>
- [51] Chou, D., & Jiang, M. (2021). A survey on data-driven network intrusion detection. *ACM Computing Surveys*, 54(9), Article 182. <https://doi.org/10.1145/3472753>.
- [52] Guo, C., Li, X., Cheng, J., Yang, S., & Gong, H. (2025). Continual Learning for Intrusion Detection Under Evolving Network Threats. *Future Internet*, 17(10), 456. <https://doi.org/10.3390/fi17100456>
- [53] Zhang, X., Zhao, R., Jiang, Z., Chen, H., Ding, Y., Ngai, E. C. H., & Yang, S.-H. (2025). Continual learning with strategic selection and forgetting for network intrusion detection. The University of Hong Kong; Southern University of Science and Technology; University of Reading. <https://arxiv.org/pdf/2412.16264>
- [54] Arreche, O., Abdallah, M. A comparative analysis of DNN-based white-box explainable AI methods in network security. *EURASIP J. on Info. Security* 2025, 16 (2025). <https://doi.org/10.1186/s13635-025-00201-x>
- [55] PapersFlow Research Team. (2026). Evaluation metrics for XAI methods: Research guide. PapersFlow. Evaluation metrics for XAI methods
- [56] Ogunseyi, T. B., Thiagarajan, G., He, H., Bist, V., & Du, Z. (2026). Performance Analysis of Explainable Deep Learning-Based Intrusion Detection Systems for IoT Networks: A Systematic Review. *Sensors*, 26(2), 363. <https://doi.org/10.3390/s26020363>
- [57] Anna L. Buczak, A. L., & Erhan Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications Surveys & Tutorials*, 18(2), 1153–1176. <https://doi.org/10.1109/COMST.2015.2494502>
- [58] Ashwin, N. (2022). A survey: Deep learning based intrusion detection systems for IoT frameworks. *International Journal of Scientific Research in Science and Technology*, 9(3), 616–623
- [59] Nguyen Truong, K., Sun, K., Wang, S., Guitton, F., & Guo, Y. (2020). Privacy Preservation in Federated Learning: An insightful survey from the GDPR perspective. *arXiv*. <https://arxiv.org/abs/2011.05411>
- [60] Zhu, L., Liu, Z., & Han, S. (2019). Deep leakage from gradients. *Advances in Neural Information Processing Systems*, 32. <https://doi.org/10.48550/arXiv.1906.08935>
- [61] Acar, A., Aksu, H., Uluagac, A. S., & Conti, M. (2018). A survey on homomorphic encryption schemes: Theory and implementation. *ACM Computing Surveys*, 51. <https://doi.org/10.48550/arXiv.1704.03578>
- [62] Li, H., & Li, Z. (2025, July 9). Homomorphic encryption in federated learning: A systematic review of current methods and open challenges. *SSRN*. <https://dx.doi.org/10.2139/ssrn.5358246>
- [63] Tan, Q., Wu, S., & Tao, Y. (2023). Privacy-Enhanced Federated Learning for Non-IID Data. *Mathematics*, 11(19), 4123. <https://doi.org/10.3390/math11194123>
- [64] European Parliament & Council of the European Union. (2016). Regulation (EU) 2016/679 (General Data Protection Regulation). Official Journal of the European Union. <https://gdpr-info.eu/>
- [65] European Parliament & Council of the European Union. (2024). Regulation laying down harmonised rules on artificial intelligence (AI Act). Official Journal of the European Union. <https://artificialintelligenceact.eu/the-act/>
- [66] European Data Protection Board (2020). Guidelines 05/2020 on consent and automated decision-making. <https://www.edpb.europa.eu>
- [67] National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). <https://nvlpubs.nist.gov/>
- [68] National Assembly of the Republic of Korea. (2011, as amended). Personal Information Protection Act. <https://elaw.klri.re.kr/>
- [69] Personal Information Protection Commission. (2021, July 9). AI personal information protection self-checklist. <https://www.pipc.go.kr>
- [70] National Intelligence Service. (2024, November 29). Korea's 2024 national cybersecurity white paper <https://www.kisa.or.kr/EN/302/form?postSeq=73>
- [71] Vijayan, J. (2021, November 30). 5 tips for reducing false positive security alerts. *CSO Online*. <https://www.csoonline.com/article/571649/5-tips-for-reducing-false-positive-security-alerts.html>