A survey of Intrusion Detection and Prevention System using Machine Learning

Nguyen Binh Thuc Tram[20520815], Nguyen Bui Kim Ngan[20520648], and Vo Anh Kiet[20520605]

UIT, VNU-HCM, Ho Chi Minh city, Vietnam
20520815@gm.uit.edu.vn, 20520648@gm.uit.edu.vn, 20520605@gm.uit.edu.vn

Abstract. In the ever-evolving landscape of cybersecurity, the imperative to safeguard digital domains against pernicious intrusion attempts stands as an immutable axiom. Central to this ongoing battle is the imperative need for robust Intrusion Detection and Prevention Systems (IDPS) that can adeptly safeguard digital assets from unauthorized access and nefarious exploits. The relentless pursuit of enhanced detection accuracy and diminished false positives and negatives in IDPS solutions has spurred a flurry of research endeavors. This pursuit is bolstered by the steadfast goal of thwarting not only known threats but also zero-day attacks. The emergence of machine learning (ML) has engendered a burgeoning array of prospects within this domain, captivating the attention of scholars and researchers alike. It has garnered substantial popularity as an efficacious means to discern and mitigate network intrusions with remarkable efficiency and heightened precision. In this survey, we present the concept of IDPS and Machine learning, subsequently venturing into the dynamic realm of ML-based IDPS solutions. This survey provides an extensive compendium of research efforts, aiming to distill their quintessence, analyze their intricacies, and discern their comparative merits.

Keywords: IDS · IPS · Machine learning

1 Introduction

Within this survey, first, we elucidate the core tenets and foundational concepts intrinsic to IDS, IPS, and machine learning (ML). As a beacon illuminating the intellectual journey, this exposition empowers readers with the necessary acumen to navigate the profound intricacies of this interdisciplinary field. Second, this study delves into the realm of contemporary research on ML-based intrusion detection and prevention systems. We conduct an exhaustive review of several recent published works, their techniques, the machine learning methodologies they deploy, the datasets instrumental to their experimentation, and key evaluation metrics employed both advantages and disadvantages. Finally, a preliminary evaluation of the amassed results will be discussed, offering a nuanced assessment of the overarching effectiveness, capabilities, and nuanced general and specific characteristics that pervade the researchs under consideration.

2 Intrusion Detection System

IDS, or "Intrusion Detection System," is composed of two words, "intrusion" and "detection system." Intrusion refers to unauthorized access to information or network systems that affects three factors: security (Confidentiality), integrity (Integrity), and availability (Availability). The detection system is a security mechanism that identifies unauthorized intrusions. Therefore, IDS is a continuous monitoring system tasked with overseeing network systems, inspecting network traffic, checking for any suspicious activities that may impact the aforementioned C-I-A factors. Figure 1 illustrates a basic network system equipped with IDS.

IDS can be classified based on the monitoring scope and implementation techniques. The classification of IDS is illustrated in Figure 2:

Based on the monitoring scope, IDS is divided into two types: Host-based IDS (HIDS) and Network-based IDS (NIDS). HIDS is deployed on hosts with the task of monitoring host activities to detect violations of policies and unusual behaviors. NIDS is implemented on network systems to protect the network system from unauthorized intrusions. NIDS continuously monitors network traffic and examines packets within the network to detect intrusions.

On the other hand, IDS based on implementation techniques is categorized into two types: Signature-based IDS (S-IDS) and Anomaly-based IDS (A-IDS).

S-IDS, also known as signature-based detection, uses known ideas and recognized patterns to identify signs of attacks. These indicators are stored in a database of signatures, and data patterns are matched against these stored indicators to detect attacks. The advantage of S-IDS is its high effectiveness in detecting pre-known attacks. However, this method lacks the ability to detect variant and new attacks because there are no patterns available in the database for these attacks. Detecting many new patterns requires a large signature database, and when an attack intends to infiltrate the system, checking patterns to determine the type of attack also consumes resources.

A-IDS, or "Anomaly-based IDS," relies on determining which operational threshold is normal and which is abnormal. Any deviation from the normal threshold is considered an anomalous behavior. The advantage of A-IDS is its ability to detect new unknown attacks and adjust normal thresholds for different networks and applications. However, the main disadvantage is the high rate of false positives because it is challenging to find the boundary between the normal and abnormal threshold to detect intrusions.

3 Intrusion Prevention System

IPS, or Intrusion Prevention System, is a key component in network security infrastructure, offering a higher level of sophistication than Intrusion Detection Systems (IDS). IPS not only monitors and detects intrusion activities and threats but also has the ability to take real-time actions in response to detected attacks across the entire network without human intervention. Depending on

pre-configured settings, IPS can issue alerts for malicious behaviors or automatically halt attacks under hardware or software devices.

The main features of IPS include the ability to automatically detect and prevent network attacks, flexibility to customize security policies according to specific network environments, and the capability to interact with other security infrastructure components such as firewalls and security management systems.

IPS can be categorized based on the timeline of attacks and action-based platforms. In terms of action-based platforms, there are Host-based Intrusion Prevention Systems (HIPS) focusing on monitoring and preventing activities on individual hosts, while Network-based Intrusion Prevention Systems (NIPS) primarily monitor network traffic and transmitted packets.

Regarding the timeline of attacks, the categorization includes distinguishing between known attacks and unknown attacks, often referred to as "zero-day attacks," by intrusion detection defense systems.

However, IPS also faces significant challenges. One of the major hurdles is the accuracy in differentiating between legitimate activities and network intrusion attempts. IPS systems must detect and prevent threats without disrupting legitimate user and application activities. The balance between preventing attacks and maintaining network system flexibility is a complex issue.

Furthermore, maintaining an updated database to identify new threats requires continuous effort and resources from security administrators. The trade-off between accuracy and system performance presents another challenge, as enhancing accuracy often results in increased latency within the system.

4 Machine Learning

This study focuses on the integration of machine learning techniques in Intrusion Detection and Prevention Systems (IDPS). Machine learning, a subset of artificial intelligence, involves the development of algorithms and models that enable computers to learn and make predictions or decisions without being explicitly programmed.

Advantages of incorporating machine learning in IDPS include enhanced detection accuracy and efficiency. Machine learning algorithms can identify intricate patterns and anomalies in network traffic, allowing for more accurate and timely intrusion detection. Additionally, machine learning can adapt and evolve over time, improving detection capabilities as it processes more data.

However, there are also some disadvantages to using machine learning in IDPS. One significant challenge is the need for substantial amounts of labeled training data, which can be both time-consuming and costly to obtain. Furthermore, machine learning models may suffer from false positives and negatives, leading to potential misclassification of normal activities as intrusions or vice versa. Additionally, the potential for adversarial attacks, where malicious actors intentionally manipulate data to deceive the machine learning models, is a concern that needs to be addressed in designing robust and resilient IDPS using machine learning.

5 Performance Evaluation

The evaluation of an Intrusion Detection and Prevention System (IDPS) using machine learning involves assessing its performance using various metrics such as True Positives (TP), True Negatives (TN), False Positives (FP), False Negatives (FN), accuracy, precision, recall, and F1 score.

True Positives (TP) represent the instances where the system correctly identifies intrusions. True Negatives (TN) denote the instances where the system correctly identifies normal activities. False Positives (FP) occur when the system mistakenly flags normal activities as intrusions. False Negatives (FN) occur when the system fails to detect actual intrusions.

The explains TP, TN, FP, and FN in a 2x2 matrix format.

Actual/Predicted	**Positive (P)**	**Negative (N)**
Positive (P)	True Positives (TP)	False Positives (FP)
Negative (N)	False Negatives (FN)	True Negatives (TN)

Table 1. Explain of True Positive, True Negative, False Positive, and False Negative

Accuracy is the proportion of correctly classified instances out of the total instances. It gives an overall measure of the system's correctness. Precision is the proportion of true positives among all instances classified as positives, providing insights into the system's ability to avoid false alarms. Recall is the proportion of true positives among all actual positives, demonstrating the system's capacity to detect intrusions. The F1 score is the harmonic mean of precision and recall, providing a balanced measure of a system's performance in terms of both false positives and false negatives. It is particularly useful when a balance between precision and recall is essential for the application.

Accuracy signifies the classifier's capacity to assess the entirety of the dataset:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Precision is the classification ability to correctly detect attacks out of the total positive predictions. It can be computed as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

Recall (Detection Rate) is the classification ability to correctly predict attacks from actual attacks. It can be computed as:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

The F1-score combines both precision and recall into a single metric, offering a balanced assessment of the model's predictive performance:

$$\text{F1-score} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{4}$$

The False Alarm Rate (FAR) is the probability of incorrectly identifying a condition or event that doesn't exist, representing the frequency of false positive detections in a system. Lowering the false alarm rate is essential for improving the accuracy and trustworthiness of detection systems by reducing instances of erroneously indicating the presence of a condition.

$$\text{FAR} = \frac{FP}{FP + TN} \tag{5}$$

Evaluating an IDPS using these metrics allows for a comprehensive assessment of its efficiency in detecting and preventing intrusions while considering the trade-off between false alarms and missed detections.

6 Literature Review

Al-Emadi et al [1] uses 3 deep learning models CNN, LSTM and GRU on the NSL-KDD dataset. This article processes data using One-hot encoding technique, normalizing values, CNN model includes 1 input layer, 1 Convolutional layer followed by 1 Max pooling layer, 1 Dense layer to connect neurons and 1 layer. Dropout avoids overfitting, and finally an output layer uses the softmax function. After experimenting with different deep learning techniques, CNN is found to have parameters of accuracy, F1 score, recall and precision that are superior to other techniques. Specifically, these parameters are all higher than 97%.

Chaabouni et al. [2] proposed an intrusion detection and prevention system for the oneM2M service layer based on machine learning. They extracted 26 features from oneM2M request-response messages to generate GFlows for machine learning. A 3-level ML approach was used, with binary classification by deep learning, semi-supervised learning for known vs unknown threats, and shallow ML like J48 for sub-type classification. The system detected oneM2M flooding, amplification, and protocol exploit attacks using a dataset of 165,253 attacks and 58,020 benign oneM2M messages. With deep learning binary classification, they achieve a false alarm rate of 13.39%, accuracy of 86.91%, precision of 86.61%, recall of 97.41%, and model size of 18KB. For flooding classification using J48, they reported accuracy of 92.32%, precision of 92.95%, recall of 93.80%, false positive rate of 1.53%, model size of 290KB, and training time of 9,280ms. The main advantages are the distributed architecture suitable for IoT, use of edge ML for efficiency, multiple levels of classification, and handling of unknown threats. Key limitations are the lack of evaluation for semi-supervised classification, need for large retraining datasets, and no real-time testing.

Krishna et al. [3] proposed an integrated intrusion detection and prevention system using deep learning. They extracted 41 features from the KDDCup99 dataset and used preprocessing like one-hot encoding and feature scaling. A Multi-Layer Perceptron (MLP) deep learning model was trained to detect DOS, Probe, R2L, and U2R attacks. The system achieves an accuracy of 91.41% with the MLP model. For intrusion prevention, they blocked malicious packets using iptables rules based on the detection results. The main advantages are high accuracy with deep learning, integration of IDS and IPS, and real-time detection. However, the evaluation is limited to just reporting accuracy, without comparing to other methods or analyzing false alarms, precision, recall, etc. Also, the system is not tested on a real network.

Wisanwanichthan et al. [4] proposed a Double-Layered Hybrid Approach (DLHA) that is better than a single ML classifier and the ensemble method. The proposed approach is composed of two layers that work in a cascading manner, where the first layer is to detect DoS and Probe, and the second layer is to detect Remote2Local (R2L) and User2Root (U2R). In their solution, they divided the NSL-KDD training dataset into two groups. The first group contains all classes and the second one contains only the U2R, R2L and normal classes in order to have a dedicated classifier for detecting rare attacks i.e., R2L and U2R amongst normal connections. This group-divided strategy allows the algorithm to focus on low-frequency attacks at the second layer. They also first adopted Intersectional Correlated Feature Selection (ICFS), in which intersecting features of different attacks against others are selected. In this process, they performed feature selection on the two groups using the Pearson Correlation Coefficient (PCC), which is used to select important features between two random variables. After the ICFS was completed, they normalized the data, performed one-hot encoding and PCA respectively. Naive Bayes (NB) is selected as a classifier for Group 1 and Support Vector Machine (SVM) is selected as a classifier for Group 2. As the dimensions are reduced in the data transform process, NBC is suitable for dealing with a large amount of connection, while to deal with linear and non-linear optimization problems, SVM creates the best hyperplane in a high-dimensional space in order to separate two classes with the maximum margin between them. They conducted experiments on the NSL-KDD dataset and the testing dataset was left unseen. The results achieved 88.97% in accuracy, 90.57% in F1 score, 88.17% in precision, and 93.11% in detection rate with 11.82% of false alarm rate. They also conducted a detailed analysis to explore the detection rates of each class, 92.4% on DoS, 90.87% on Probe, 96.67% on R2L, and 100% on U2R.

Bo Cao et al [5] used ADRDB (a combination of two algorithms Adaptive Synthetic Sampling (ADASYN) and Repeated Edited nearest neighbors (RENN)) on 3 data sets UNS_NB15, NSL-KDD and CIC- IDS2017 to create samples to minimize the quantity difference between positive and negative samples. Next, the system uses the Random Forest algorithm combined with Pearson correlation analysis (RFP algorithm) to select features to solve the problem of feature redundancy. The CNN model will consist of a Convolutional layer fol-

lowed by a layer combining Averagepooling and Maxpooling to extract spatial features. The GRU deep learning model is used to extract long-distance dependent information features. Finally, there is a final output layer that uses the softmax function for classification. The proposed intrusion detection model is evaluated based on the UNSW_NB15, NSL-KDD, and CIC-IDS2017 datasets, and the experimental results show that the classification accuracy reaches 86.25%, 99.69%, 99.65%, which are 1.95%, 0.47% and 0.12% higher than that of the same type of CNN-GRU, and can solve the problems of low classification accuracy and class imbalance well.

Studies	**Year**	**Feature Processing Techniques**	**ML methods**	**Attacks Detected**
[1]	2020	Proposed method	CNN	U2R, R2L
[2]	2020	GFlows	J48	oneM2M flooding, amplification and protocol exploits
[3]	2020	N/A	Multi-Layer Perceptron (MLP)	DOS, Probe, R2L, U2R
[4]	2021	ICFS and PCA	NB and SVM	DoS, Probe, Remote2Local (R2L), User2Root (U2R)
[5]	2022	Random Forest + Pearson correlation analysis	CNN - GRU	Normal, Dos, Probe, R2L, U2R
[6]	2023	ExtraTreeClassifier	RF, LR, KNN, ANN, NB	DDoS

Table 2. Summary of Literature Review: Features

Sadhwani et al. [6] presented the concept of lightweight Internet of Things (IoT) networks, comprising devices characterized by limited computational resources, including reduced battery life, processing capabilities, memory, and notably, minimal security provisions, rendering them susceptible to Distributed Denial of Service (DDoS) attacks and malware infiltration. Their research was primarily centered around the development of a DDoS attack detection model, aimed at discerning between legitimate and malicious network traffic patterns through the identification of anomalies. The handling of missing values, data standardization using Standard Scalar, feature selection using ExtraTreeClassifier wherein only the 15 best features are extracted, and anomaly detection using a classifier were performed. The evaluation encompassed the assessment of five ML algorithms, namely Logistic Regression (LF), Random Forest (RF), Naive Bayes (NB), Artificial Neural Network (ANN), and K Nearest Neighbor (KNN). Experimental analysis was performed on the TON-IOT and BOT-IOT datasets, addressing binary and multiple-class classification scenarios, which cor-

respond to DDoS attacks and all attacks, respectively. In the TON-IOT dataset and its multiple-class classification task utilizing 15 features, RF exhibits superior performance with a remarkable accuracy rate of 100%. Conversely, in binary classification within the same dataset, NB emerges as the top-performing algorithm, also achieving a flawless accuracy score of 100%. Similarly, in the BOT-IOT dataset, both multiple-class and binary classifications benefit from the effectiveness of NB, delivering an accuracy rate of 100% in each scenario.

Studies	**Dataset**	**False Alarm Rate**	**Accuracy**	**Precision**	**F1 score**	**Recall**
[1]	NSL-KDD	N/A	97,01	100,00	98,48	97,01
[2]	oneM2M	1.53	92.32	92.95	N/A	93.80
[3]	KDDCup99	N/A	DOS 91.41 Probe 90.56 R2L 91.32 U2R 90.39	N/A	N/A	N/A
[4]	NSL-KDD	11.82	88.97	88.17	90.57	93.11
[5]	UNSW_NB15,	N/A	86.25	86.92	86.25	86.59
	NSL-KDD		99.69	99.65	99.69	99.70
	CIC-IDS2017		99.65	99.63	99.65	99.64
[6]	TON-IOT	0	100	100	100	100
	BOT-IOT	0	100	100	100	100

Table 3. Summary of Literature Review: Evaluation Metrics.

7 Evaluation

The evaluated intrusion detection and prevention systems offer diverse methodologies and performances. The approach leveraged CNN, LSTM, and GRU models, identifying CNN as the most effective with remarkable accuracy, F1 score, recall, and precision, all-surpassing 97%. Another system employed a layered approach, integrating deep learning for binary classification and semi-supervised learning, effectively distinguishing specific attack types. In a different system, an integrated approach achieved high accuracy using an MLP model, although a more comprehensive evaluation is necessary for a broader comparison and a detailed analysis of additional metrics.

Furthermore, the Hybrid approach showcased competitive accuracy and detection rates, particularly excelling in identifying low-frequency attacks. The utilization of advanced techniques, such as ADRDB and CNN-GRU, successfully addressed class imbalance and boosted classification accuracy. Additionally, in the context of lightweight IoT networks, another system achieved notable accuracy rates in DDoS detection using various machine learning algorithms. However, to comprehensively assess their efficiency and applicability in real-world

intrusion detection scenarios, a broader evaluation encompassing comparison with other state-of-the-art methods and a wider range of performance metrics is essential.

Studies	Advantages	Disadvantages
[1]	Proposed solution for overfitting problem and CNN model results in high Percision	Use the Old-fashioned dataset is NSL-KDD
[2]	Distributed architecture suitable for IoT, edge ML for efficiency, multiple levels of classification and handles unknown threats	No results reported for semi-supervised classifier, large dataset needed for retraining and no real-time evaluation
[3]	High accuracy, real-time detection, integrated IDS and IPS	Limited evaluation - only accuracy reported and no comparison with other methods
[4]	Their solution achieved outstanding accuracy for R2L and U2R with respective accuracy of 96.67% and 100%	The accuracy for large classes (DoS, Pobe) was not as good as the other efficient solutions, high FAR
[5]	Strong feature extractioncapability, high detection accuracy and low false alarm rate when dealing with large-scalehigh-dimensional network data	Not perform a good job in dimensionality large in the dataset leads to significantly reduced training speed.
[6]	Their solution achieved outstanding results (using NB) against DDoS attacks in IoT networks.	The solution only focused on DDoS and did not provide details about the different types of other attacks. They just generally said all attacks, which had quite poor evaluation metrics results or required high training/prediction time

Table 4. Summary of Literature Review: advantages and disadvantages

8 Conclusions

This survey provides an overview of the features and operation of intrusion detection systems (IDSs) and intrusion prevention systems (IPSs), and highlights the advantages of using machine learning technology in the attack detection process. The papers also demonstrate the potential of applying machine learning technology to IDSs and IPSs, achieving many impressive performance metrics in a variety of network attacks. However, these models still have some drawbacks, such as: low performance when applied in practice, requiring a large dataset and taking a long time to train, and not yet being able to simultaneously detect multiple types of network attacks, and false positives still occur.

Through the synthesis of recent research, the team believes that machine learning methods are a promising solution for the mission of ensuring network security for businesses. However, the main goal of building an AI-IDS system is its ability to accurately predict attacks that surpass human analytical capabilities in order to help analysts better process security events and unknown attacks; therefore, building an AI-IDS system to analyze real-time network traffic becomes a very worthy research topic. In the future, the team will explore unsupervised learning algorithms to generate new datasets. Additionally, the team will explore how to optimize neural network models to help improve model performance and save time and costs.

References

1. Sara Al-Emadi, Aisha Al-Mohannadi, and Felwa Al-Senaid. Using deep learning techniques for network intrusion detection. In *2020 IEEE international conference on informatics, IoT, and enabling technologies (ICIoT)*, pages 171–176. IEEE, 2020.
2. Nadia Chaabouni, Mohamed Mosbah, Akka Zemmari, and Cyrille Sauvignac. A onem2m intrusion detection and prevention system based on edge machine learning. In *NOMS 2020-2020 IEEE/IFIP Network Operations and Management Symposium*, pages 1–7. IEEE, 2020.
3. Akhil Krishna, Ashik Lal, Athul Joe Mathewkutty, Dhanya Sarah Jacob, and M Hari. Intrusion detection and prevention system using deep learning. In *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pages 273–278. IEEE, 2020.
4. Treepop Wisanwanichthan and Mason Thammawichai. A double-layered hybrid approach for network intrusion detection system using combined naive bayes and svm. *IEEE Access*, 9:138432–138450, 2021.
5. Bo Cao, Chenghai Li, Yafei Song, Yueyi Qin, and Chen Chen. Network intrusion detection model based on cnn and gru. *Applied Sciences*, 12(9):4184, 2022.
6. Sapna Sadhwani, Baranidharan Manibalan, Raja Muthalagu, and Pranav Pawar. A lightweight model for ddos attack detection using machine learning techniques. *Applied Sciences*, 13(17):9937, 2023.