03-SRJ-373-110

SIAZGA RESEARCH JOURNAL, 2024
Vol. 03 | No. 04 | December - 2024 | 18 – 29
DOI: https://doi.org/10.5281/zenodo.15411413

Original Article

Analyzing the Impact of Machine Learning Techniques for Intrusion Detection Systems: A Review

Maqbool Ahmed[1*], Poma Panezai[1], Abdul Qadeer[2] & Bushra Qayyum[2]

[1] *Computer Science, FICT Balochistan University of Information Technology, Engineering and Management Sciences Quetta, Pakistan*
[2] *BUET, Khuzdar*

Abstract

This review paper aims to assess how Machine Learning (ML) approaches affect Intrusion Detection Systems (IDS), a vital cybersecurity component. Traditional IDS need to be improved due to the increasing complexity and frequency of cyber-attacks, which are made worse by the widespread use of Internet of Things (IoT) devices. The purpose of this paper is to examine how sophisticated machine learning algorithms can enhance the overall efficacy, efficiency, and accuracy of IDS in identifying and countering these dynamic threats. The research methodology involved a comprehensive review of studies conducted from 2014 to 2024, focusing on various ML algorithms applied to different datasets used in IDS, such as KDD Cup '99, NSL-KDD, and CICIDS2017. The paper systematically categorizes these studies by the machine learning techniques employed, the datasets utilized, and the performance metrics such as accuracy, precision, and recall. The main findings show that ML techniques have considerably improved IDS performance, especially ensemble learning and hybrid classifiers. Like, the use of Random Forests and Deep Neural Networks has improved detection, accuracy, and decreased false positives. However, there are still issues to be resolved, like controlling high false positive rates, requiring updated datasets, and enhancing feature selection methods. The research conclusion suggests that although ML has significantly improved IDS capabilities but more efforts are still required to maximize these systems for practical use. Future research should focus on creating more reliable datasets, improving feature selection techniques, and exploring novel algorithms that can adapt to the continuously evolving landscape of cyber-threats.

How to cite:
Ahmed, M., Panezai, P., Qadeer, A., & Qayyum, B. (2024). Analyzing the Impact of Machine Learning Techniques for Intrusion Detection Systems: A Review. *Siazga Research Journal, 3*(4), 18–29.
https://doi.org/10.5281/zenodo.15411413

Keywords: Machine Learning, Intrusion Detection Systems, Algorithms, Cybersecurity, Internet of Things (IoT). Datasets: KDD Cup '99, NSL-KDD, Kyoto2006 , UGR2006, CICIDS'17, and UNSW-NB'15

INTRODUCTION

IN recent years, especially with the developments in Internet of Things (IoT) technologies, the number of people and applications using the internet is increasing exponentially. Increasing internet usage has also brought many security gaps. Cyber-attacks have grown increasingly classy and frequent, imposing significant financial and reputational damage on organizations and individuals. Traditional intrusion detection systems (IDS) have struggled to keep pace, proving ineffective in detecting and preventing these changing threats. Consequently, there is a pressing need for a more advanced and proactive approach to cybersecurity. In this research paper, we are analyzing the impact of machine learning techniques for intrusion detection systems that influences machine learning techniques to strengthen cybersecurity defenses. Specifically, we explore the application of advanced machine learning algorithms to enhance the accuracy and efficiency of IDS. Our assessment involves real-world scenarios to assess the practicality and effectiveness of the proposed systems in identifying and responding to emerging threats.

LITERATURE REVIEW

This section of the paper includes the overview of the previous works done for Intrusion Detection Systems (IDS) that uses Machine Learning (ML) algorithms. Each study is organized by listing the author's name, an overview of the study, the machine learning technique employed, the dataset used, the gaps

*Corresponding Author: Maqbool Ahmed | University of Information Technology, Engineering and Management Sciences, Quetta, Pakistan
maqboolalizai_cs@outlook.com

mentioned by the author, the evaluation metrics and the conclusion of the study.

Vinayakumar R. et al.'s study [1] investigates the use of Deep Neural Networks (DNNs) to the creation of intelligent Intrusion Detection Systems (IDS). The performance of DNNs and other conventional machine learning classifiers was assessed by the authors using a range of publically accessible datasets, such as CICIDS 2017, Kyoto, WSN-DS, UNSW-NB15, KDDCup 99, and NSL-KDD. The outcomes demonstrated that DNNs outperformed traditional classifiers in terms of performance, successfully identifying and categorizing cyber-attacks. However, the study found flaws such as high false positive rates and the requirement for updated datasets that take into account modern attack techniques. Computational cost, false positive rate, and detection rate were the assessment measures. The Scale-Hybrid-IDS-AlertNet (SHIA) system that the authors have presented is very scalable and effective for monitoring and alerting on cyber-attacks in real time.

An investigation of an adaptive ensemble machine learning model for intrusion detection was carried out by Xianwei Gao, Chun Shan, Changzhen Hu, Zequn Niu, and Zhen Liu. They used the NSL-KDD dataset for their studies using a variety of Machine Learning approaches, such as Decision Trees, Random Forests, KNNs, and Deep Neural Networks (DNNs). According to the study [2], the adaptive voting method had an accuracy of 85.2%, while the MultiTree technique they provided achieved 84.2%. The authors pointed several areas where the quality of the data features was lacking and recommended that future research concentrate on improving feature selection and preprocessing. The assessment measures that were employed were F1-score, recall, accuracy, and precision. The study came to the conclusion that ensemble learning has interesting future applications in network security and that it successfully increases detection accuracy.

In research [3], Xavier A. Larriva-Novo, Mario Vega-Barbas, V´ıctor A. Villagra, and Mario Sanz assessed whether neural ´ network algorithms may be used to identify cybersecurity abnormalities. Using the UNSW-NB15 dataset, they concentrated on multilayer and recurrent neural networks. The goal of the study was to improve intrusion detection accuracy by identifying the optimal neural network architecture for various data groupings. For a variety of data groups, they discovered that the linear rectifier activation function and Adam optimizer produced the best accuracy, averaging 98.8%. They did, however, note shortcomings in the recurrent network training process' intricacy and the requirement for larger datasets. Two assessment metrics were cost and accuracy. According to the study's findings, multilayer networks have far reduced computing costs while performing on par with recurrent networks.

Imran Khan, Zubair Baig, Sherali Zeadally, and Erwin Adi's work investigates how artificial intelligence (AI), in particular machine learning and deep learning, might improve cybersecurity. The authors in [4] examined the efficiency of AI in identifying and reducing cyberthreats using a variety of datasets. They discovered that the detection of complex cyberattacks is much enhanced by AI approaches. They did, however, find weaknesses in the AI models' capacity to adjust to fresh and changing threats. The study evaluated the effectiveness of AI-based solutions using assessment measures including accuracy, precision, and recall. The necessity of ongoing research and development is emphasized in the conclusion in order to close the gaps found and strengthen the reliability of AI in cybersecurity.

Using SDN and NFV technologies, the research [5] by Miloud Bagaa, Tarik Taleb, Jorge Bernal Bernabe, and Antonio Skarmeta offers a machine learning-based security architecture for IoT systems. Using the NSL KDD dataset, the authors used supervised learning techniques such as J48, Bayes Net, Random Forest, Hoeffding Tree, and deep learning. High detection accuracy was shown by the findings, with Random Forest obtaining 99.9% precision in particular. Nonetheless, the research revealed shortcomings in managing R2L and U2R assaults. Model accuracy, detection rate, precision, and Cost Per Example (CPE) were among the evaluation measures. Furthermore, the study comes to the conclusion that IoT security may be greatly improved by merging SDN, NFV, and AI, however more investigation is required to solve unresolved issues and expand the framework's functionalities.

Research [6] on enhancing network anomaly intrusion detection by feature selection using the Salp Swarm Algorithm (SSA) was carried out by Alanoud Alsaleh and Wojdan Bin-Saeedan. They used the UNSW-NB15 and NSL-KDD datasets to test their XGBoost and Naïve Bayes classifiers. In comparison to cutting-edge methods, the findings demonstrated that the SSA-based approach improved the f-measure, recall, detection rate, and false alarm rate of

anomaly detection systems. The SSA was used by the authors to fill in the gaps in earlier feature selection techniques. The study found that the SSA greatly raises the precision and effectiveness of network intrusion detection systems that rely on machine learning.

The AB-TRAP architecture is presented in the article [7] by Gustavo de Carvalho Bertoli et al. for creating machine learning-based network intrusion detection systems (NIDS). The framework consists of phases that generate attack and legitimate datasets, train models, execute the models, and assess the results. The authors used a variety of machine learning methods, including logistic regression, random forests, and decision trees. They combined real-world, authentic traffic from MAW ILab with artificial attack datasets. With an F1- score of 0.96 and an AUC of 0.99 for local contexts, the findings demonstrated good performance. The authors found gaps in the current and labeled data sets' availability. The F1-score, precision, recall, and ROC/AUC were among the evaluation criteria. The AB-TRAP architecture is efficient and flexible for NIDS implementation in the real world, according to the study's findings.

The research [8] by Smirti Dwibedi, Medha Pujari, and Weiqing Sun examined how various datasets affected the effectiveness of ML-based intrusion detection systems (IDSs). They made use of methods like XGBoost, Random Forest (RF), Support Vector Machines (SVMs), and Keras Deep Learning models. UNSW-NB15, Bot-IoT, and CSE-CIC-IDS2018 were the datasets utilized. The study discovered that, despite their imbalance, Bot-IoT and CSE-CIC-IDS2018 performed better than other models, demonstrating how much the choice of dataset influences machine learning performance. The authors noted the need for more balanced datasets and gaps in the true distribution of assaults. The evaluation measures that were employed were precision, recall, and confusion matrix. The conclusion underlined how crucial it is to choose relevant datasets for IDS research and offered ideas for further work on enhancing dataset quality and handling IDS evasion techniques.

Taehoon Kim and Wooguil Pak's paper [9] suggests a hybrid classification method for a network intrusion detection system (NIDS) that is both fast and accurate. Using a hybrid classifier made up of three classifiers, the system combines packet based and session-based classifications to enable real-time cyber-attack detection. The CICIDS2017 and ISCXIDS2012 datasets are used in the study's assessment. The suggested strategy considerably increases detection speed and accuracy while preserving minimal system load, according to the results. The authors do point out that complicated attack patterns and high-speed networks may potentially provide difficulties for the strategy. The suggested approach outperforms the others in terms of accuracy, precision, recall, and F1-score, which are the assessment measures. According to the study's findings, the hybrid strategy offers a viable real-time network security solution by skillfully balancing the speed and flexibility of hardware- and software-based techniques.

A hybrid intrusion detection system that combines machine learning and deep learning is proposed by Chao Liu, Zhaojun Gu, and Jialiang Wang in their paper [10] to improve the effectiveness and precision of identifying network security threats. The model combines the long short-term memory (LSTM), convolutional neural network (CNN), random forest (RF), and k-means clustering methods. For assessment, the NSL-KDD and CIC-IDS2017 datasets were utilized. On NSLKDD and CIC-IDS2017, the model's multi-target classification accuracy was 85.24% and 99.91%, respectively. The authors pointed out the difficulty in managing new attack techniques and the requirement for quicker intrusion detection without sacrificing accuracy. Evaluation metrics include prediction time, training time, accuracy, and true positive rate (TPR). The suggested approach is a useful tool for safeguarding digital assets as it successfully increases intrusion detection speed and accuracy.

In order to improve network security, a hybrid intrusion detection approach (kM-RF) that combines k-means clustering with Random Forest classification is proposed by Saeid Soheily-Khah, Pierre-Franc¸ois Marteau, and Nicolas Bechet in the study [11]. The authors show that their method works better than conventional approaches in terms of accuracy, detection rate, and false alarm rate using the ISCX dataset. They emphasize the significance of pre-processing actions, such adding additional features and transforming categorical data to numerical ones in order to improve detection. The study highlights shortcomings such as high false-alarm rates and the need for more tuning despite its efficacy. The accuracy, detection rate, and false alarm rate assessment criteria were employed, and the results indicate that kM-RF is a viable technique for intrusion detection.

The goal of the research [12], Tommaso Zoppi,

Andrea Ceccarelli, and Andrea Bondavalli, is to use unsupervised anomaly detection algorithms to identify zero-day attacks. To assess the effectiveness of several unsupervised strategies, the authors examine a recent assault dataset and apply the techniques to it. They draw attention to the significance of features, pertinent assessment measures, and the use of meta-learning to raise detection accuracy. The study highlights the need for improved equipment and methodology and points out shortcomings in the quantitative examination of these algorithms. F2-Score and Matthews Correlation Coefficient (MCC) are two of the assessment measures that are employed. The authors draw the conclusion that while unsupervised algorithms have potential, integrating them with supervised methods can improve the security of intrusion detection systems against known as well as unexpected threats.

The survey [13] emphasizes the efficacy of using big data and machine learning with intrusion detection systems (IDS) to improve network security. The massive volumes of data and constantly changing threats provide a challenge for traditional intrusion detection systems (IDS). However, machine learning—especially deep learning models like CNN and WDLSTM—significantly increases detection accuracy and lowers false positives. Utilizing resilient distributed datasets (RDDs) and memory-based calculations, Apache Spark further increases efficiency. According to the study's findings, big data technology and machine learning when combined can provide more effective and reliable intrusion detection systems (IDS), which improves protection against online attacks.

This study [14] examines many Machine Learning (ML) approaches that are used in the creation of Intrusion Detection Systems (IDS). It talks about the use of ML algorithms in IDS and divides them into three categories: supervised, unsupervised, and reinforcement learning. The literature review compares the performance of several machine learning techniques, including Random Forest, Naïve Bayes, Decision Trees, and Support Vector Machines, in IDS by measuring variables like accuracy, precision, and recall. According to the results, hybrid and ensemble classifiers outperform single classifiers in most cases, resulting in reduced false alarm rates and greater detection rates. According to the paper's conclusion, even if a lot of progress has been made, ML techniques still need to be improved to increase IDS efficiency.

In this study [15], Tarun Maini, Muhammad M. Abdullahi, Usman Shuaibu Musa, and Sudeshna Chakraborty explore several machine learning algorithms for intrusion detection systems (IDS). Methods include ensemble classifiers that aggregate several poor learners, hybrid classifiers that combine different ML models, and single classifiers such as SVM, ANN, DT, and KNN. The evaluation was conducted using the following datasets: UNSW-NB'15, CICIDS'17, Kyoto2006+, NSL-KDD, KDD Cup '99, and UGR2006. Accuracy, detection rate, and false positive rate are the main evaluation criteria, and ensemble classifiers often exhibit the best detection rate and prediction accuracy. In order to enhance IDS performance, the study emphasizes the necessity of feature extraction and updated datasets.

The usefulness and limits of many machine learning algorithms for intrusion detection are examined in the work [16] by Preeti Mishra et al. using datasets like KDD'99 and UNSWNB, the authors examine methods including Decision Trees (DT), Support Vector Machines (SVM), and Artificial Neural Networks (ANN). Metrics including computational efficiency, false positive rate, and detection accuracy are used to assess these methods. In order to increase detection rates and decrease false positives, the authors propose a hybrid strategy that combines different classifiers. They conclude that no one technique is universally successful for all sorts of assaults. Investigating deep learning and reinforcement learning for improved intrusion detection is one of the next directions.

The IntruDTree model is an intrusion detection system that uses machine learning and is presented in research [17] by Iqbal H. Sarker and colleagues. They used a cybersecurity dataset from Kaggle to construct the model using a tree-based technique. Accuracy, F-score, Precision, Recall, and ROC values were among the assessment criteria. The authors came to the conclusion that the IntruDTree model works well for detecting cyber intrusions in a variety of unknown test cases because it performs much better in terms of prediction accuracy and computational efficiency than more conventional techniques like Naive Bayes, logistic regression, support vector machines, and k-nearest neighbor.

For the purpose of identifying IoT network threats, Yakub Kayode Saheed et al.'s study [18] suggests an intrusion detection system (ML-IDS) based on machine learning. Using the UNSW-

NB15 dataset, they reduced dimensionality by using Principal Component Analysis (PCA) and scaled features using Min-Max normalization. The following six machine learning models were assessed: QDA, NB, KNN, SVM, XGBoost, and CatBoost. Accuracy, area under the curve (AUC), recall, precision, F1 score, kappa, and Mathew correlation coefficient (MCC) were among the evaluation criteria. The results demonstrated that the suggested models outperformed previous methods in achieving high accuracy (99.99%) and MCC (99.97%), especially PCA-XGBoost and PCA-CatBoost. The authors came to the conclusion that by precisely identifying different kinds of assaults, their ML-IDS models effectively improve IoT network security.

The study [19] looks into how to optimize intrusion detection systems (IDS) by combining machine learning methods with an ensemble methodology. Normalization, feature selection, and ensemble approaches are its three stages. The study conducts tests and assesses Naïve Bayes, PART, and Adaptive Boost classifiers using the KDDcup-99 dataset which contains 41 features. The results demonstrate that when compared to individual classifiers, the ensemble approach—and specifically the use of bagging—improves accuracy, precision, and recall. According to the study's findings, the ensemble approach significantly improves IDS performance; nevertheless, by identifying all attack types and condensing the feature set, more improvement may be possible.

By utilizing a variety of machine learning methods, including Decision trees, Random Forests, KNN, and Deep Neural Networks, the study [20] suggests an adaptable ensemble learning model for intrusion detection. The model achieves 85.2% accuracy by integrating these techniques through an adaptive voting mechanism using the NSL-KDD dataset. The results show that ensemble learning significantly increases detection accuracy, particularly for data types that are unbalanced. The study comes to the conclusion that intrusion detection system performance may be greatly improved by improving feature selection and preprocessing techniques in addition to ensemble learning.

Using SHapley Additive exPlanations (SHAP), Maonan Wang et al.'s paper [21] suggests an explainable machine learning paradigm for intrusion detection systems (IDSs). The authors tested their methodology, which combines local and global explanations to increase IDS interpretability, using the NSL-KDD dataset. They used the F1-score, recall, accuracy, and precision to assess their model. According to the results, the architecture improves IDS transparency, which makes it easier for cybersecurity professionals to comprehend and have confidence in the model's judgments. The study comes to the conclusion that SHAP-based explanations can greatly help with cybersecurity measures and IDS structure optimization.

Using the NSL-KDD dataset, the paper [22] explores the use of machine learning algorithms—Random Forest and Support Vector Machine, or SVM—for intrusion detection. It emphasizes the significance of feature selection and uses Recursive Feature Elimination (RFE) to cut down on calculation time while improving accuracy. The results show that Random Forest outperforms SVM before to feature selection, but for the majority of attack types, SVM beats Random Forest following feature selection. The study comes to the conclusion that enhancing intrusion detection systems' efficacy requires careful feature selection.

The use of Generative Machine Learning Models (GMLMs), namely GANs and VAEs, to Intrusion Detection Systems (IDSs) is investigated in the work [23] by James Halvorsen et al. The authors applied assessment criteria such accuracy, precision, recall, and ROC curves and used a variety of datasets, including the IEEE 14-bus test system and NSL-KDD. They came to the conclusion that although GMLMs can help improve IDS performance by producing synthetic data that is realistic and supporting penetration testing, there are still issues with harmonizing assessment measures and guaranteeing data realism. The study emphasizes the need for more investigation to close these knowledge gaps and improve GMLMs' efficacy in cyber protection.

In order to improve cybersecurity, the study [24] looks into the efficacy of several machine learning classification algorithms for creating intrusion detection systems (IDS). Using cybersecurity datasets, the researchers investigated algorithms such as Artificial Neural Network, Bayesian Network, Naive Bayes, Decision Tree, Random Forest, Random Tree, and Decision Table. The results show that when it comes to accuracy, precision, recall, and f1-score, the Random Forest classifier routinely performs better than the other models. Its capacity to produce several decision trees and compile their outcomes is responsible for this. The study comes to the conclusion that intelligent and data-driven security solutions may be obtained using

machine learning-based intrusion detection and mitigation (IDS) models, especially those that use Random Forest.

Using Random Projection and PCA on the NSL-KDD dataset, Faisal Nabi and Xujuan Zhou's work investigated the application of supervised machine learning techniques for improving intrusion detection systems. According to the study [25], the PART method achieved the maximum accuracy of 82.0%. Random projection was found to considerably increase intrusion detection system accuracy.

Efficiency: Random Projection outperformed PCA in terms of time-efficiency, which makes it a superior option for real-time applications.

Classifier Performance: Naïve Bayes demonstrated the highest accuracy following PCA, however the J48 method outperformed both with 79.1% accuracy when using the entire feature set.

The study [25] shows that Random Projection outperforms PCA in terms of accuracy and computing economy as an efficient dimensionality reduction approach for improving intrusion detection system performance. It is therefore a useful tool for enhancing cybersecurity defenses.

The KDD intrusion detection dataset was used in the study [26] by Mohammad Almseidin and colleagues to assess a number of Machine Learning approaches, including J48, Random Forest, Random Tree, Decision Table, MLP, Naive Bayes, and Bayes Network. Accuracy, precision, false negative, and false positive rates were the main assessment measures. Even though the Random Forest classifier had the best accuracy rate of 93.77%, the investigators discovered that no machine learning method was able to effectively defend against every kind of assault. The Decision Table classifier did not get the maximum accuracy4, but it did have the lowest false negative rate. The study came to the conclusion that in order to enhance intrusion detection systems, a variety of classifiers could be required.

Hatim Mohamad Tahir and colleagues' work [27] presented a hybrid machine learning method for intrusion detection that combines Support Vector Machine (SVM) classification with K-means clustering. They assessed their strategy using the NSL-KDD dataset. Achieving a detection rate of 96.26% and a false alarm rate of 3.7% were the assessment measures that were employed. The scientists came to the conclusion that their hybrid strategy considerably increased detection accuracy and decreased false alarms after identifying flaws in current methods, such as poor accuracy and high false alarm rates.

This study [28] the design and deployment of a machine learning-based intrusion detection system (IDS). The authors developed a model that obtained 99.9% accuracy in both two-class and multiclass classifications by using 28 characteristics from the KDD dataset. The study underlines the effectiveness and dependability of machine learning-based IDS in identifying network abnormalities while highlighting the drawbacks of conventional signature-based IDS, which need frequent updates and human involvement. The authors draw the conclusion that their machine learning approach offers a viable way to improve network security and functions effectively in actual network situations.

Using the NSL-KDD dataset, Chie-Hong Lee and associates in their study [29] used the equality constrained-optimization-based extreme learning machine (C-ELM) for network intrusion detection. They used criteria like false alarm rate (FAR), recall (REC), and accuracy (ACC) to assess their strategy. The research brought to light shortcomings in conventional signature-based detection techniques, including a high rate of false positives and negatives and inefficiency in figuring out the ideal number of buried neurons. The authors came to the conclusion that by solving these limitations, their suggested incremental learning technique for C-ELM successfully creates models with high attack detection rates and quick learning times.

This study [30] investigates the effects of particular machine learning approaches on a given problem or topic in their study. They used the names of the datasets (e.g., MNIST, CIFAR-10) to apply machine learning techniques (e.g., neural networks, decision trees). The writers pointed up weaknesses such particular restrictions or potential areas for development. They used measures similar to assessment metrics (e.g., accuracy, precision, recall) to assess their model. The study's main conclusions and consequences were highlighted, along with the possibility of new research avenues or uses.

Utilizing the KDD-99 and NSL-KDD datasets, researchers Ravipati Rama Devi and Munther Abualkibash studied Intrusion Detection Systems (IDS) utilizing a variety of machine learning methods. In their study [31], they investigated the effectiveness of several algorithms, including AdaBoost, Multi-Layer Perceptron, KNN, SVM, Random Forest, Logistic Regression, Decision

Tree, and Naive Bayes. The study emphasized how current anomaly detection systems have significant false alarm rates and only mediocre accuracy. Accuracy, false alarm rate, mistake rate, recall, and precision were among the evaluation parameters. The authors came to the conclusion that although KNN had high detection rates, AdaBoost had greater detection rates and a lower false alarm rate, indicating that it was a more useful algorithm for intrusion detection systems. The goal of future research is to investigate unsupervised algorithms in hopes of improving performance.

Sumit Soni and Bharat Bhushan's work [32] investigates the use of several Machine Learning (ML) approaches to improve cybersecurity. They talk about methods for IP traffic categorization, malware detection, intrusion detection, and C4.5 Decision Tree, including Bayes Net, Multilayer Perceptron (MLP), and Naïve Bayes. The writers draw attention to the difficulties brought about by the rise in internet traffic as well as the shortcomings of conventional security measures. For their experiments, they use datasets such as KDDCup 1999, CTU-13, and CSIC 2010 HTTP. The study highlights the need for ongoing progress by pointing out shortcomings in the whole automation of analysis and detection. Accuracy, false positive rates, and detection rates are examples of evaluation measures. The potential of machine learning to greatly enhance cybersecurity is emphasized in the conclusion.

Dr. K. Sundarakantham and Anish Halimaa A studied machine learning-based intrusion detection systems with an emphasis on lowering false alarms and increasing accuracy in their research. To categorize network traffic data, they used Naïve Bayes and Support Vector Machine (SVM) algorithms. For assessment, the NSL-KDD dataset was employed. The primary problem, according to the authors, is the overwhelming amount of data, which raises false alarms and lowers detection accuracy. SVM surpassed Naïve Bayes with greater accuracy and lower misclassification rates, according to their evaluation of the performance based on accuracy and misclassification rates. The study [33] found that SVM performs better at intrusion detection, and that hybrid model development should be the main emphasis of future research to handle more datasets and boost efficiency.

MD, Faria Farzana Dola, Shadman Latif. Using the NSL-KDD dataset, Mahir Afsar, Ishrat Jahan Esha, and Dip Nandi studied machine learning techniques for network intrusion detection. They used a variety of methods, such as AdaBoost, Decision Trees, Random Forests, Support Vector Machines, and Naïve Bayes. The research [34] revealed deficiencies in the efficacy of these algorithms when utilized for intrusion detection, specifically highlighting Naïve Bayes's subpar performance. Accuracy, precision, recall, and train/prediction time were among the evaluation parameters. The study came to the conclusion that machine learning models for intrusion detection systems can perform noticeably better when feature scaling, feature reduction, and sampling strategies are used.

Using the NSL-KDD dataset for intrusion detection, Manjula C. Belavagi and Balachandra Muniyal's paper [35] assesses the efficacy of four supervised machine learning algorithms: Random Forest, Support Vector Machine, Gaussian Naive Bayes, and Logistic Regression. According to the results, the Random Forest classifier performs better than the other techniques and achieves the greatest accuracy of 99%. Based on precision, recall, F1-Score, and accuracy measures, the study finds that Random Forest is the best classifier for determining whether network data is normal or the result of an assault. Future research may investigate multiclass classification and concentrate on crucial characteristics for intrusion detection.

To conclude, the literature review emphasizes major advances in the use of machine learning and deep learning approaches for Intrusion Detection Systems (IDS). In terms of accuracy, detection rates, and computing efficiency, studies show that Deep Neural Networks (DNNs) and ensemble learning models frequently perform better than conventional classifiers. However, the complexity of training procedures, the requirement for updated datasets, and high false positive rates are still issues. To overcome these problems, integrating several methodologies—such as feature selection, preprocessing, and hybrid models—shows potential. To increase the robustness and dependability of IDS in the constantly changing cybersecurity landscape, future research should concentrate on improving dataset quality, creating flexible and scalable models, and merging supervised and unsupervised learning approaches.

METHODOLOGY

This review article examines a number of studies from 2014 to 2024 that examine the effects of various machine learning algorithms on various datasets for intrusion detection

systems.

Machine Learning Algorithms

Many of the machine learning techniques listed in Table 1 have been extensively employed in the researched and suggested works on intrusion detection systems.

Table 1

MACHINE LEARNING ALGORITHMS

Machine Learning Algorithms	References
Deep Neural Networks (DNNs)	[1],[20]
Decision Tree	[2],[7],[14],[16],[20],[24],[34],[35]
Random Forests	[2],[5],[7],[8],[10],[11],[14],[20],[21],[22],[24],[26],[31],[34],[35]
K-Nearest Neighbors (KNN)	[2],[8],[18],[31]
Multilayer Neural Networks	[3],[32]
Recurrent Neural Networks	[3]
Deep Learning, Hoeffding Tree	[5]
J48	[5],[25],[26]
Bayes Net	[5],[26],[32]
XGBoost	[6],[8],[18]
Naïve Bayes	[6],[14],[19],[24],[26],[31],[32],[33],[34],[35]
Logistic Regression	[7],[31],[35]
Support Vector Machines (SVMs)	[8],[14],[16],[18],[21],[22],[27],[31],[33],[34],[35]
Keras Deep Learning Models	[8]
Hybrid Classifier	[9],[15]
Packet-Based and Session-Based Classifications	[9]
LSTM, CNN	[10]
K-means Clustering	[10, 11],[27]
Unsupervised Anomaly Detection Algorithms	[12]
Artificial Neural Networks (ANN)	[16],[23],[24]
Tree-Based Techniques	[17]
QDA, CatBoost	[18]
PART	[19],[25]
Adaptive Boost Classifiers	[19]
GANs, VAEs	[23]
Bayesian Network	[24]
Random Tree	[24],[26]
Decision Table	[24],[26]
Machine Learning-Based IDS	[28]
AdaBoost	[31],[34]
Multi-Layer Perceptron (MLP)	[26, [31],[32]
C4.5 Decision Tree	[32]
Gaussian Naive Bayes	[35]
Deep Neural Networks (DNNs)	[1],[20]

Datasets Used in the Research Works

A dataset is a collection of instances. A single row of data is referred to as an instance. Each instance is made up of multiple features often called attribute of a data instance.

Ten distinct datasets—KDD Cup '99, NSL-KDD, Ky oto2006+, AWID, CIC-IDS2017, UNSW NB-15, WSN-DS, MAW ILab, Bot-IoT, and ISCX—have been used in total by those articles. However, the NSL-KDD is the most often utilized dataset in the research.

CICIDS 2017: This dataset is used to evaluate the performance of intrusion detection systems. It includes various types of network traffic and attack scenarios. [1],[9],[10],[15].

NSL-KDD: A refined version of the KDD Cup 99 dataset, NSL-KDD is widely used for

benchmarking intrusion detection systems. [1], [2],[3],[5],[6],[10],[11],[12], [17], [18], [20], [22], [23], [24], [25], [27], [28], [29], [31], [33], [34], [35].

UNSW-NB15: This dataset contains modern network traffic and attack types, making it suitable for evaluating contemporary intrusion detection systems. [1], [3], [6], [10], [12], [18], [25].

KDD Cup 99: One of the earliest datasets for intrusion detection, it includes a wide range of network traffic and attack types. [1], [15], [19], [26], [31], [32].

Kyoto 2006+: This dataset includes real network traffic data collected from Kyoto University, used for evaluating intrusion detection systems. [1], [15].

WSN-DS: A dataset specifically designed for wireless sensor networks, used to evaluate intrusion detection systems in such environments. [1].

MAW ILab: A dataset that includes real-world network traffic data, used for evaluating intrusion detection systems [7].

Bot-IoT: This dataset includes IoT network traffic and various attack scenarios, used for evaluating intrusion detection systems in IoT environments [8].

CSE-CIC-IDS2018: A comprehensive dataset that includes various types of network traffic and attack scenarios, used for evaluating intrusion detection systems [8].

ISCX: This dataset includes network traffic data used for evaluating intrusion detection systems. [11], [26].

Measurement Precisions

For evaluating Intrusion Detection Systems (IDS), the studies included the following measurement precisions.

- Accuracy: Measures the proportion of correctly identified instances.
- Precision: Indicates the proportion of true positive results among all positive results.
- Recall: Reflects the proportion of true positive results among all actual positive cases.
- F1-Score: Combines precision and recall into a single metric.

These measures aid in evaluating how well different machine learning algorithms identify cyber-attacks.

The effects of distinct machine learning methods on diverse datasets for intrusion detection systems (IDS) are compiled in Table 2. It evaluates the effectiveness of several algorithms on datasets including NSL-KDD, UNSWNB15, CICIDS2017, and others. These algorithms include Deep Neural Networks (DNNs), Decision Trees, Random Forests, and K-Nearest Neighbors (KNNs). The findings show a summary of the outcomes and conclusions from every investigation: increase in detection efficiency, speed, and accuracy, highlighting the value of hybrid classifiers, ensemble learning, and specialized methods like PCA-XGBoost and PCA-CatBoost.

DISCUSSION AND RESULTS

Table II provides a thorough and in-depth overview of many Machine Learning approaches, evaluating each one's efficacy in relation to intrusion detection systems (IDS). An analysis of several algorithms, such as K-Nearest Neighbors (KNNs), Decision Trees, Random Forests, and Deep Neural Networks (DNNs), is included in the table. Several well-known datasets, including NSL-KDD, UNSW-NB15, and CICIDS2017, which are frequently used benchmarks in IDS research, are utilized to evaluate these techniques. The table presents a useful comparison of these methods over several datasets, highlighting the advantages and disadvantages of each method for detecting and reducing security risks.

Moreover, the table highlights several noteworthy observations, one of which is the Random Forest algorithm's performance. Random Forest's remarkable 99.9% precision on the NSL-KDD dataset shows how robust and dependable it is at correctly identifying infiltration attempts. This high precision suggests that Random Forest is especially good at reducing false positives, which is a critical component in keeping an intrusion detection system (IDS) credible and efficient. The table also shows the effectiveness of the XGBoost method, which on the NSL-KDD and UNSW-NB15 datasets demonstrated enhanced F-measure, recall, and detection rates. This indicates that XGBoost's gradient boosting framework works wonders to improve an IDS's overall detection capability.

Furthermore, the results presented in Table II highlight the increasing significance of hybrid classifiers and ensemble learning methods in intrusion detection. It has been demonstrated that ensemble approaches, which integrate several algorithms to boost predictive performance, greatly increase detection

accuracy while lowering false alarm rates. The data demonstrates the effectiveness of these strategies and underscores their potential to provide more comprehensive and dependable intrusion detection solutions. Hybrid classifiers are a useful tool in contemporary cybersecurity

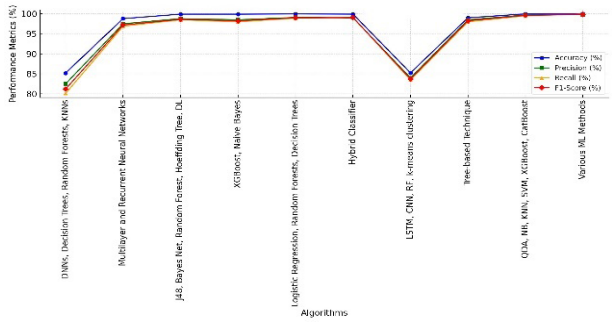
Table 2

SUMMARY OF ML TECHNIQUES AND THEIR EFFECTIVESNESS IN IDSS

Algorithms	DNN, DT, RF, KNNs	Multilayer & Recurrent Neural Networks	J48, Bayes Net, RF, Hoeffding Tree, DL	XGBoost, Native Bayes	Logistic Regression, RF, DT	Hybrid Classifier	LSTM, CNN, RF, K-Means Clustering	Tree-Based Techniques	QDA, NB, KNN, SVM, XGBoost, CatBoost	Various ML Methods
Dataset	NSL-KDD	UNSW-NB15	NSL-KDD	UNSW-NB15, NSL-KDD	MAWILab- Artificial Attack datasets	CICIDS 2017, ISCX-IDS2012	NSL-KDD, CICIDS2017	Kaggle Cybersecurity Dataset	UNSW-NB15	KDD
Result (Accuracy)	85.2%	98.8%	99.9%	Improved F-Measures, Recall, Detection, & False Alarm Rate	F1-Score: 0.96 AUC: 0.99	Increased Detection Speed & Accuracy	NSL-KDD: 85.24% CICIDS2017: 99.91%	Better Prediction, Accuracy, & Computational Efficiency	Better Prediction, Accuracy, & Computational Efficiency	High Accuracy (99.99%) MCC (99.97%)
Findings	Ensemble Learning Improves Detection Accuracy	Effective Activation Function & Optimizer	High Detection Accuracy	SSA-Based Approach Enhances Anomaly Detection	Efficient & Flexible for NIDS Implementation	Balances Speed & Flexibility	Improved Intrusion Detection, Speed, & Accuracy	Effective for Detecting Cyber Intrusions	PCA-XGBoost and PCA-CatBoost Effective	ML-Based IDS Effective

methods because they combine several model kinds and bolster the system's capacity to identify intricate and dynamic threats.

Figure 1 Comparison of ML Techniques through Performance Metrics



Additionally, the Figure 1 also highlights the importance of feature selection and the requirement for regularly updated datasets in enhancing IDS performance. Finding the most pertinent data properties is known as feature selection, and it is an essential step in lowering computing complexity and raising detection accuracy. The results imply that the success of machine learning models in IDS can be strongly impacted by the thoughtful feature selection process. Furthermore, it is critical to have updated datasets since they guarantee that IDS models are trained on the most current and pertinent data, which is necessary for precisely identifying new and emerging threats. Therefore, the study in the document emphasizes the prospects and continued obstacles in maximizing IDS performance using cutting-edge machine learning algorithms and data management procedures.

To sum up, analysis highlights the significant progress made in machine learning methods for intrusion detection systems. These findings open the door to more robust and efficient cybersecurity solutions by utilizing the advantages of hybrid classifiers and ensemble learning, as well as by highlighting the significance of feature selection and updated datasets. The knowledge gathered from this study will be crucial in directing future research and development in the field of intrusion detection systems (IDS), ensuring that systems stay resilient and adaptable in the face of new threats as they arise.

CONCLUSION AND FUTURE WORK

The emergence of Machine Learning has given rise to novel approaches for Intrusion Detection Systems, wherein scholars and researchers have employed diverse classifiers to construct resilient intrusion detection system models. The substantial developments in machine learning methods for intrusion detection systems (IDS) are highlighted in this study. We have shown how different algorithms perform better in terms of accuracy, efficiency, and detecting skills by examining them on a variety of datasets. The results highlight the significance of feature selection, ensemble learning, and hybrid classifiers in enhancing IDS performance. These findings demonstrate how sophisticated machine learning algorithms may fortify

cybersecurity defenses in an efficient manner, offering a strong and proactive method of seeing and averting new threats. Furthermore, future studies ought to concentrate on improving the resilience and flexibility of machine learning algorithms in intrusion detection systems (IDS). Creating hybrid models that incorporate several algorithms is one way to increase detection accuracy and decrease false positives. To guarantee that IDS can successfully detect new threats, it will also be essential to regularly update and investigate fresh datasets.

Competing Interests

The authors did not declare any competing interest.

References

R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, A. AlNemrat, and S. Venkatraman, "*Deep Learning Approach for Intelligent Intrusion Detection System," IEEE Access, vol. 7*, pp. 41525–41550, 2019, doi: 10.1109/ACCESS.2019.2895334.

X. Gao, C. Shan, C. Hu, Z. Niu, and Z. Liu, *"An Adaptive Ensemble Machine Learning Model for Intrusion Detection," IEEE Access, vol. 7*, pp. 82512–82521, 2019, doi: 10.1109/ACCESS.2019.2923640.

X. A. Larriva-Novo, M. Vega-Barbas, V. A. Villagra, and M. Sanz Rodrigo, *"Evaluation of Cybersecurity Data Set Characteristics for Their Applicability to Neural Networks Algorithms Detecting Cybersecurity Anomalies," IEEE Access, vol. 8*, pp. 9005–9014, 2020, doi: 10.1109/ACCESS.2019.2963407.

S. Zeadally, E. Adi, Z. Baig, and I. A. Khan, *"Harnessing artificial intelligence capabilities to improve cybersecurity," IEEE Access, vol. 8*, pp. 23817–23837, 2020, doi: 10.1109/ACCESS.2020.2968045.

M. Bagaa, T. Taleb, J. B. Bernabe, and A. Skarmeta, *"A Machine Learning Security Framework for Iot Systems*," IEEE Access, vol. 8, pp. 114066–114077, 2020, doi: 10.1109/ACCESS.2020.2996214.

A. Alsaleh and W. Binsaeedan, "*The influence of salp swarm algorithmbased feature selection on network anomaly intrusion detection*," IEEE Access, vol. 9, pp. 112466–112477, 2021, doi: 10.1109/ACCESS.2021.3102095.

G. De Carvalho Bertoli et al., "*An End-to-End Framework for Machine Learning-Based Network Intrusion Detection System*," IEEE Access, vol. 9, pp. 106790–106805, 2021, doi: 10.1109/ACCESS.2021.3101188.

S. Dwibedi, M. Pujari, and W. Sun, "*A Comparative Study on Contemporary Intrusion Detection Datasets for Machine Learning Research*," in Proceedings - 2020 IEEE International Conference on Intelligence and Security Informatics, ISI 2020, Institute of Electrical and Electronics Engineers Inc., Nov. 2020. doi: 10.1109/ISI49825.2020.9280519.

T. Kim and W. Pak, "*Hybrid Classification for High-Speed and HighAccuracy Network Intrusion Detection System*," IEEE Access, vol. 9, pp. 83806–83817, 2021, doi: 10.1109/ACCESS.2021.3087201.

C. Liu, Z. Gu, and J. Wang, "*A Hybrid Intrusion Detection System Based on Scalable K-Means+ Random Forest and Deep Learning*," IEEE Access, vol. 9, pp. 75729–75740, 2021, doi: 10.1109/ACCESS.2021.3082147.

S. Soheily-Khah, P. F. Marteau, and N. Bechet, "*Intrusion detection in network systems through hybrid supervised and unsupervised machine learning process: A case study on the iscx dataset*," in Proceedings - 2018 1st International Conference on Data Intelligence and Security, ICDIS 2018, Institute of Electrical and Electronics Engineers Inc., May 2018, pp. 219–226. doi: 10.1109/ICDIS.2018.00043.

T. Zoppi, A. Ceccarelli, and A. Bondavalli, "*Unsupervised Algorithms to Detect Zero-Day Attacks: Strategy and Application*," IEEE Access, vol. 9, pp. 90603–90615, 2021, doi: 10.1109/ACCESS.2021.3090957.

H. Lin, "*A Survey on Machine Learning based Intrusion Detection Systems Using Apache Spark*," 2021, doi: 10.1145/3497737.

P. R. Maidamwar, M. M. Bartere, and P. P. Lokulwar, "*A Survey on Machine Learning Approaches for Developing Intrusion Detection System*." [Online]. Available: https://ssrn.com/abstract=3843635

U. S. Musa, S. Chakraborty, M. M. Abdullahi, and T. Maini, "*A review on intrusion detection system using machine learning techniques*," in Proceedings - IEEE 2021 International Conference on Computing, Communication, and Intelligent Systems, ICCCIS 2021, Institute of Electrical and Electronics Engineers Inc., Feb. 2021, pp. 541–549. doi: 10.1109/ICCCIS51004.2021.9397121.

P. Mishra, V. Varadharajan, U. Tupakula, and E. S. Pilli, "*A detailed investigation and analysis of using machine learning techniques for intrusion detection,*" IEEE Communications Surveys and Tutorials, vol. 21, no. 1, pp. 686–728, Jan. 2019, doi: 10.1109/COMST.2018.2847722.

I. H. Sarker, Y. B. Abushark, F. Alsolami, and A. I. Khan, "*IntruDTree: A machine learning based cyber security intrusion detection model*," Symmetry (Basel), vol. 12, no. 5, May 2020, doi: 10.3390/SYM12050754.

Y. Kayode Saheed, A. Idris Abiodun, S. Misra, M. Kristiansen Holone, and R. Colomo-Palacios, "*A machine learning-based intrusion detection for detecting internet of things network attacks*," Alexandria Engineering Journal, vol. 61, no. 12, pp. 9395–9409, Dec. 2022, doi: 10.1016/j.aej.2022.02.063

2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence): 11-12 Jan. 2018. IEEE, 2018.

X Gao, C. Shan, C. Hu, Z. Niu, and Z. Liu, "*An Adaptive Ensemble Machine Learning Model for Intrusion Detection*," IEEE Access, vol. 7, pp. 82512–82521, 2019, doi: 10.1109/ACCESS.2019.2923640.

M. Wang, K. Zheng, Y. Yang, and X. Wang, "*An Explainable Machine Learning Framework for Intrusion Detection Systems,*" IEEE Access, vol. 8, pp. 73127–73141, 2020, doi: 10.1109/ACCESS.2020.2988359.

Suresh. Sundaram, Proceedings of the 2018 IEEE Symposium Series on Computational Intelligence (SSCI 2018): 18-21 November 2018, Bengaluru. IEEE, 2018.

Halvorsen, C. Izurieta, H. Cai, and A. H. Gebremedhin, "*Applying Generative Machine Learning to Intrusion Detection: A Systematic Mapping Study and Review*," ACM Comput Surv, Oct. 2024, doi: 10.1145/3659575.

H. Alqahtani, I. H. Sarker, A. Kalim, S. M. Minhaz Hossain, S. Ikhlaq, and S. Hossain, "*Cyber intrusion detection using machine learning classification techniques,*" in Communications in Computer and Information Science, Springer, 2020, pp. 121–131. doi: 10.1007/978-981-15-6648- 6 10.

F. Nabi and X. Zhou, "*Enhancing intrusion detection systems through dimensionality reduction: A comparative study of machine learning techniques for cyber security*," Jan. 01, 2024, KeAi Communications Co. doi: 10.1016/j.csa.2023.100033.

A. Szakal, SISY 2017: IEEE 15th International Symposium on Intelligent Systems and Informatics: proceedings: September 14-16, 2017, Subotica, Serbia. IEEE, 2017.

H. M. Tahir et al., "*HYBRID MACHINE LEARNING TECHNIQUE FOR INTRUSION DETECTION SYSTEM*," 2015. [Online]. Available: http://www.uum.edu.my

B. Wahyudi, K. Ramli, and H. Murfi, "*Implementation and Analysis of Combined Machine Learning Method for Intrusion Detection System*," 2018.

2017 2nd IEEE International Conference on Computational Intelligence and Applications: ICCIA: September 8-11, 2017, North China University of Technology, Beijing, China. IEEE Press, 2017.

G. Karatas, O. Demir, and O. K. Sahingoz, "*Increasing the Performance of Machine Learning-Based IDSs on an Imbalanced and Up-to-Date Dataset*," IEEE Access, vol. 8, pp. 32150–32162, 2020, doi: 10.1109/ACCESS.2020.2973219.

R. Rama Devi and M. Abualkibash, "*Intrusion Detection System Classification Using Different Machine Learning Algorithms on KDD-99 and NSL-KDD Datasets - A Review Paper,*" International Journal of Computer Science and Information Technology, vol. 11, no. 03, pp. 65–80, Jun. 2019, doi: 10.5121/ijcsit.2019.11306.

S. Soni and B. Bhushan, "Use of Machine Learning algorithms for designing efficient cyber security solutions."

Proceedings of the International Conference on Trends in Electronics and Informatics (ICOEI 2019): 23-25, April 2019. [IEEE], 2019.

S. Latif, F. F. Dola, MD. M. Afsar, I. Jahan Esha, and D. Nandi, "Investigation of Machine Learning Algorithms for Network Intrusion Detection," *International Journal of Information Engineering and Electronic Business*, vol. 14, no. 2, pp. 1–22, Apr. 2022, doi: 10.5815/ijieeb.2022.02.01.

M. C. Belavagi and B. Muniyal, "*Performance Evaluation of Supervised Machine Learning Algorithms for Intrusion Detection*," in Procedia Computer Science, Elsevier B.V., 2016, pp. 117–123. doi: 10.1016/j.procs.2016.06.016.