Journal of Information Systems Engineering and Management
2025, 10(15s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Ensemble Learning Classifiers and Hybrid Feature Selection for Enhancing Intrusion Detection System Performance

Vishwas Sharma1, Dharmesh J. Shah2

[1] Sankalchand Patel University, Visnagar, Gujarat, India

2 Indrashil University, Kadi, Gujarat, India

ARTICLE INFO

Received: 15 Dec 2024
Revised: 25 Jan 2025
Accepted: 14 Feb 2025

ABSTRACT

Network environments must be protected against a variety of cyber-attacks using IDS. The advancement of ML methodologies has yielded notable improvements in intrusion detection system (IDS) capabilities, including greater real-time analysis, adaptability, and accuracy of detection. This study uses machine learning algorithms to give an analytical comparison of several IDS models. The research covers a variety of machine learning approaches, such as supervised and hybrid strategies. We assess these models' performance using important measures including computational efficiency, precision, recall etc. The results show that although supervised machine learning models provides high accuracy, but when used in hybrid model including Random Forests and SVM improves performance. The result is a hybrid model that leverages the strengths of each approach. For instance, Random Forest can provide a robust feature representation, while SVM can refine the decision boundary, leading to a more accurate and reliable classification model. This combination often yields better performance than using any single algorithm alone.

Keywords: Cyberattack, Cybersecurity, Intrusion Detection Systems, Machine Learning, network traffic, supervised machine learning, SVM.

INTRODUCTION

In the digital era, robust defences against cyberattacks are now required to safeguard networked systems. Identification and mitigation of any security breaches are crucial tasks for intrusion detection systems, or IDS. Traditional IDS approaches, primarily based on signature detection, have struggled to keep pace with the rapidly evolving threat landscape. Consequently, there has been a paradigm shift towards incorporating machine learning (ML) techniques in IDS to enhance their efficacy. Advanced IDS can be developed by machine learning, which has the potential to recognize patterns and learn from data. By leveraging ML, IDS can achieve higher detection rates, reduce false positives, and adapt to emerging threats. This has led to extensive research and development of various ML-based IDS models, each with its own strengths and weaknesses.

This study aims to provide an analytical evaluation of several IDS models that rely on machine learning. We will investigate ensemble techniques like RF and Gradient Boosting, as well as supervised machine learning techniques like SVM and Decision Trees. Our goal in examining these measures is to shed light on how well each machine learning method performs when it comes to intrusion detection. In addition, we will talk about the usefulness of using these models in actual network contexts.

Cybersecurity has become a major worry for businesses, governments, and individuals in today's linked world. Threats to digital infrastructures are growing in sophistication and ubiquity along with them. Through the detection and mitigation of possible security breaches, IDS, are essential to the protection of these infrastructures. However, the constant development and improvement of IDS technology is required due to the swift evolution of cyber threats. The dynamic nature of contemporary cyberattacks makes traditional IDS methods, which are frequently based on signature detection and rule-based systems, difficult to keep up with. These conventional systems are limited by their reliance on predefined patterns and can be easily circumvented by novel or polymorphic threats. Consequently, the need for more adaptive and intelligent IDS solutions has never been greater. The growing intricacy and refinement of cyber-attacks presents formidable obstacles for traditional IDS. Traditional IDS, primarily reliant on signature-based and rule-based methodologies, are often inadequate in detecting novel and evolving threats. This limitation highlights the need for more sophisticated, flexible, and perceptive intrusion detection techniques. A promising path to improving IDS's capabilities is through ML. IDS can identify patterns in past data, learn from them, and instantly adjust to new threats by utilizing machine learning algorithms. This change from static to dynamic threat detection could greatly increase IDS's efficiency and accuracy by lowering false positives and uncovering hitherto undiscovered attack pathways.[1]

OBJECTIVES

This research study aims to perform a thorough analytical evaluation of different machine learning approaches used in IDS. In order to determine the best practices for boosting IDS capabilities, this study will assess and compare the performance of several ML algorithms. Provided a Comparative Analysis: To conduct a thorough comparative analysis that highlights the strengths and weaknesses of each ML approach, using multiple benchmark datasets to ensure a robust evaluation across diverse conditions and attack scenarios.

LITERATURE SURVEY

Comparative studies evaluating different ML techniques for IDS are relatively sparse but crucial for identifying the most effective approaches. Recent advancements in deep learning have significantly impacted IDS. Meidan et al. (2022) demonstrated the effectiveness of CNNs for IDS by leveraging their capability to automatically extract features from network traffic data. Their work showed that CNNs could outperform traditional machine learning algorithms in detecting complex attack patterns due to their powerful feature extraction abilities [2]. Similarly, Vanlalruata et al. (2023) explored the use of LSTM networks for time-series data in network intrusion detection. They highlighted that LSTMs are particularly adept at capturing temporal dependencies in network traffic, which is crucial for detecting sequential attack patterns and anomalies [3]. Their results indicated that LSTMs could enhance the detection of sophisticated attacks that involve temporal sequences. Xiaoning et al. (2023) proposed an ensemble learning approach combining multiple classifiers such as RF, SVM, and DT. Their study emphasized that ensemble methods leverage the strengths of various classifiers to improve overall detection performance and reduce false positives and negatives [4]. The integration of diverse classifiers results in a more robust IDS capable of addressing a wide range of attack types. The Author Hasan et al. (2022) introduced a hybrid model that integrates deep learning and traditional machine learning techniques [5]. Hybrid models represent a promising direction for enhancing IDS performance by leveraging the benefits of multiple methodologies. Additionally, Henry et al. (2023) investigated the use of autoencoders to identify deviations from normal behaviour, a key aspect of detecting novel attacks. Their approach demonstrated that autoencoders could effectively learn data representations and identify outliers, making them suitable for detecting unknown or emerging threats [6]. Whereas Yang et al. (2023) examined the use of clustering algorithms such as DBSCAN and K-means for intrusion detection in large-scale networks. Their research showed that clustering can effectively group similar attack patterns, aiding in the detection of previously unseen attacks and improving the overall robustness of IDS [7]. Another author Abiodun et al. (2022) provided a comprehensive review of advanced feature extraction techniques. Their findings emphasize that well-designed feature extraction methods can significantly boost the performance of IDS by providing more relevant and discriminative features [8]. While Jaw et al. (2023) explored the use of genetic algorithms for selecting relevant features in IDS. Their study demonstrated that genetic algorithms could effectively reduce dimensionality and enhance model efficiency by identifying and retaining the most informative features [9]. In addition, author Alduailij et al. (2023) conducted a comparative study on performance metrics for IDS, in this research they highlighted the selection of appropriate metrics of IDS models. They proposed a framework for assessing IDS performance that considers various aspects of detection capabilities and model reliability [10]. Here author Al Lail et al. (2023) reviewed recent datasets such as UNSW-NB15 and CICIDS, noting their relevance for contemporary IDS research. Their review emphasized the importance of using updated datasets that reflect current network conditions and attack vectors for accurate model evaluation [11]. The author Siamak et al. (2024) investigated how federated learning might be used in distributed network systems. Their work addressed issues with data sharing and security by demonstrating how federated learning permits cooperative model training over numerous nodes while maintaining data privacy [12]. While here author Zhao et al. (2023) examined methods for enhancing the interpretability of ML models in IDS. Their research emphasized the importance of making IDS decisions transparent and understandable to users, which is crucial for trust and effective decision-making [13].

METHODS

The proposed model is a hybrid classifier with combined RF and SVM models for distinguishing between normal and attack instances in a dataset can potentially leverage the strengths of both algorithms. This combination can be implemented using techniques like stacking or voting.

Here's a step-by-step guide to implementing a hybrid RF and SVM algorithm for binary classification (normal vs. attack):

Proposed hybrid Algorithm:

Start

The size of the Input dataset for training is $N \times M$

The size of the Input dataset for testing is $N \times M$

where: N represents attacks number and M represents selected features number

At the output:

Correctly detected data and incorrectly detected data are classified.

Efficiency of the algorithm for classification

Begin

Step 1: Load and Preprocess Data

a. Load Dataset: Import the training and testing datasets.

b. Preprocess Data: Handle missing values, scale features, and ensure labels are formatted correctly.

Step 2: Split Data

a. Split Dataset: Divide the dataset into training and testing sets.

Step 3: Initialize Base Models

a. Initialize Models: Set up Random Forest and SVM classifiers.

Step 4: Choose Hybrid Approach

a. Select Method: Decide between stacking or voting to combine the base models.

Step 5: Train Hybrid Model

a. Train Models: Fit the selected hybrid model to the training data.

Step 6: Make Predictions

a. Predict: Use the trained model to generate predictions on the test set.

Step 7: Evaluate Model

a. Calculate Performance Metrics:

Correctly detected data

Incorrectly detected data

Step 8: End

This flow provides a high-level overview of the steps involved in implementing a hybrid RF and SVM model for classification. It helps to visualize the workflow and understand how different components fit together.

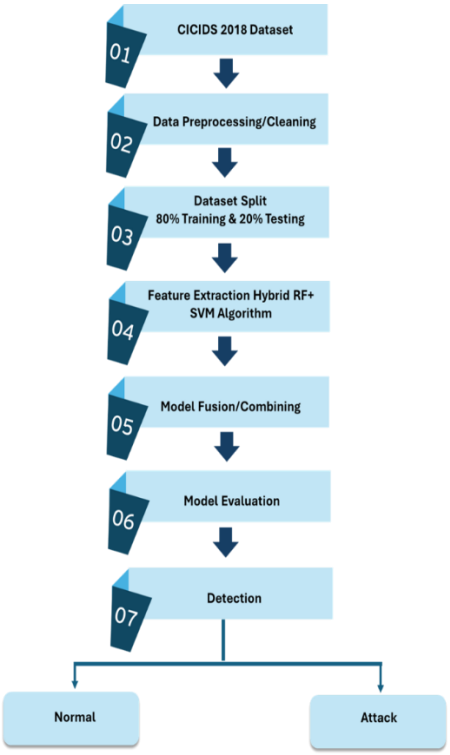

Figure 1. Proposed Model

RESULTS

The results of the experiments are presented in tabular form, comparing the performance of different ML models across the selected datasets. The experimental results demonstrate the effectiveness of various ML models in IDS applications. Deep learning models, consistently outperformed traditional supervised and unsupervised learning methods across all datasets. Ensemble methods also showed robust performance, indicating their potential for practical deployment in IDS.

These findings provide a comprehensive understanding of the strengths and limitations of different ML approaches for IDS, guiding future research and practical implementations to enhance the security of digital infrastructures. The dataset contains NULL values. Then, pre-processing is applied to the CICDS2018 datasets, and continuous NULL values are removed. To eliminate the NULL values, the row is deleted from the dataset. The percentage of malware-type samples against benign-type samples in the CICIDS2018 collection is out of balance. One may argue that there are significantly fewer samples of the malware type than those of the benign type. The unbalanced CICIDS2018 data is transformed into balanced data for binary classifiers using the SMOTE Tomek approach.

Table 2: Performance Comparison of ML Models on CICIDS2018 Dataset

Model	Accuracy	Precision	Recall	F1 Score	Execution Time (s)
KNN	96.6%	93.4%	95.7%	96.8%	2.18 secs
SVM	69.9%	54.6%	92.8%	60.3%	47.09 secs
CART	97.0%	97.6%	92.0%	96.5%	0.74 secs
RF	97.1%	99.1%	90.7%	96.5%	0.66 secs
ABoost	97.5%	96.5%	95.0%	97.1%	12.92 secs
LR	96.0%	97.1%	89.2%	95.2%	5.87 secs
NB	74.0%	53.9%	79.8%	77.5%	0.19 secs
LDA	93.2%	90.9%	86.3%	94.0%	0.68 secs
QDA	84.8%	71.8%	59.8%	94.9%	0.37 secs
MLP	94.4%	88.8%	91.4%	96.1%	23.48 secs
Proposed Model	98.98%	97.9%	97.6%	99.9%	49.88 secs

Results Analysis

The training dataset has unbalanced classes, which leads to unbalanced learning. Semantic problems include unbalanced classification. While there is some variation in the unbalanced class distribution, more specialized strategies may be needed for modelling significantly unbalanced data. Binary Classification: To divide items in a given collection into two categories, binary, or binomial, classification methods are applied. The IDS dataset contains the binary classification of a target as either benign or malicious. The dataset's aim is unbalanced. Imblearn's dataset is balanced using the SMOTE Tomek approach. amalgamate the library. Therefore, the Random Over Sampler function provided by the imblearn. oversampling module is utilized to balance the dataset. The optimum parameter for each classifier is found using the hyper-parameter approaches, namely Grid Search CV and Randomized Search CV, based on the dataset. A binary-class classifier is used to classify the target in the dataset. Thus, based on binary-class classifiers, ten widely-used machine-learning classification models are employed. The performance of these models is evaluated using following metrics, such as F1_score, Accuracy, Precision, Recall, and Total time (in seconds) for each approach.

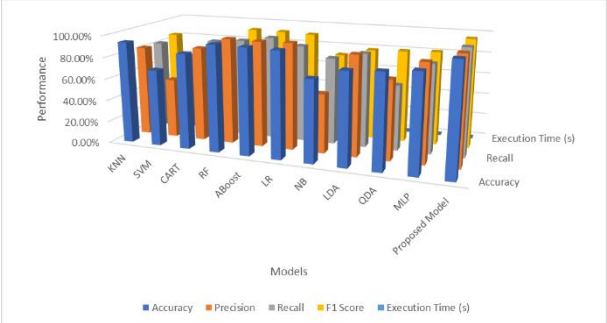

Figure 2. Performance comparison of Models

A testing model's accuracy is determined by how effectively it can differentiate between benign and malicious.

DISCUSSION

An extensive analytical comparison of several ML techniques used with IDS was offered in this research study. We have discovered important information about the effectiveness, drawbacks, and performance of a wide range of ML models by assessing them on several benchmark datasets using ensemble methods, supervised learning, unsupervised learning, and deep learning techniques.

This study emphasizes how machine learning approaches can greatly improve the capabilities of intrusion detection systems. We have produced important insights that can direct future research and real-world deployments by methodically comparing a range of machine learning models. Sustaining resilient and adaptable cybersecurity defences will depend on continuous innovation and improvement in ML-based IDS.

REFRENCES

[1] Alsaedi, M., Hussain, M., & Saeed, F. (2020). Enhanced IDS using deep learning techniques for IoT applications. IEEE Access, 8, 157387-157396. DOI: 10.1109/ACCESS.2020.3018472

[2] Meidan, Y., Bohadana, M., & Breitenstein, A.: Anomaly-based network intrusion detection using autoencoders. IEEE Transactions on Information Forensics and Security, 17, 1128-1139 (2022).

[3] Vanlalruata H., Jamal H.: An Intelligent Intrusion Detection System Using Convolutional Neural Networks. Telematics and Informatics Reports (11) 100077(2023).

[4] Xiaoning W., Jia L., Chunjiong Z.: Network intrusion detection based on multi-domain data and ensemble-bidirectional LSTM. EURASIP Journal on Information Security (2023) 5 (2023).

[5] Hasan, T., Malik, J., Bibi, I., Khan, W. U., Al-Wesabi, F. N., Dev, K.; Securing industrial Internet of Things against botnet attacks using hybrid deep learning approach, IEEE Trans. Network. Science and Engineering, vol. 10, no. 5, pp. 2952-2963, (2023).

[6] Henry, A., Gautam, S., Khanna, S., Rabie, K., Shongwe, T., Bhattacharya, P., Chowdhury, S.: Composition of hybrid deep learning model and feature optimization for intrusion detection system. Sensors 23(2), 890 (2023).

[7] Yang, L., Song, Y., Gao, S., Hu, A., Xiao, B.: Griffin: real-time network intrusion detection system via ensemble of autoencoder in SDN. IEEE Trans. Network. Serv. Manage. 19(3), 2269–2281 (2022).

[8] Abiodun M., Absalom E., Laith A., Belal A., Jia H.: K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. Information Sciences (622), 178-210 (2023).

[9] Jaw, E.; Wang, X.: Feature Selection and Ensemble-Based Intrusion Detection System: An Efficient and Comprehensive Approach. Symmetry (13), 1764 (2021).

[10] Alduailij, M., Khan, Q.W., Tahir, M., Sardaraz, M., Alduailij, M., Malik, F.: Machine-Learning-Based DDoS Attack Detection Using Mutual Information and Random Forest Feature Importance Method. Symmetry (14), 1095(2022).

[11] Al Lail, M., Garcia, A. O., livo, S.: Machine Learning for Network Intrusion Detection—A Comparative Study. Future Internet (15), 243 (2023).

[12] Siamak, L., Marcus, G., Marius, P.: Benchmarking the benchmark—Comparing synthetic and real-world Network IDS datasets. Journal of Information Security and Applications (80), 103689 (2024).

[13] Zhao, Z., Yong, Z., Hao, L., Shenbo, L., Wei, C., Zhigang, Z., Lijun, T.: Federated continual representation learning for evolutionary distributed intrusion detection in Industrial Internet of Things. Engineering Applications of Artificial Intelligence (135), 108826 (2024).