Archives available at journals.mriindia.com

International Journal on Advanced Computer Theory and Engineering

ISSN: 2319-2526
Volume 14 Issue 01, 2025

Designing an Efficient Machine Learning-Based Intrusion Detection System for Enhanced Cyber Security in Modern Networks

Winit N. Anandpwar[1], Shweta M. Barhate[2], Mahendra P. Dhore[3]
[1]*Research Scholar, Dr. Ambedkar College, Nagpur, RTM Nagpur University, Nagpur, Maharashtra, winit.anand@gmail.com*
ORCID iD: 0009-0001-5758-2190
[2]*Associate Professor, Department of Electronics and Computer Science, PGTD, R.T.M. Nagpur University Nagpur Maharashtra India , shwetab73@yahoo.com*
[3]*Professor & Pro-Vice Chancellor, Sant Gadge Baba Amravati University, Amravati, Maharashtra, India, mpdhore@rediffmail.com*

Peer Review Information

Submission: 17 Jan 2025
Revision: 14 Feb 2025
Acceptance: 15 March 2025

Keywords

Intrusion Detection System
Machine Learning
Cybersecurity
Random Forest
Support Vector Machine
Network Security

Abstract

To protect important data and assets in the digital age, strong protection is a must. Intrusion detection systems (IDS) have become one of the most important ways to protect against online dangers. But because cyberattacks are getting smarter and more complicated, IDS systems need to be more efficient and effective. This study shows an effective way to improve attack detection in current networks using machine learning. Random Forest (RF) and Support Vector Machines (SVM) are two cutting-edge machine learning methods that are used in the suggested method. Because they can handle large, complicated datasets and give correct classification results, these algorithms were picked. The method involves preparing data from network traffic by pulling out traits that can be used to spot bad behaviour. Both algorithms are taught on a labelled dataset that includes both regular and attack traffic. This lets them learn trends that can point to hacking attempts. Metrics like accuracy, precision, recall, and F1-score are used to compare how well the two methods work. Traditional signature-based IDS don't work as well as our suggested method, which gets higher detection rates and lower false positive rates. Using machine learning methods, especially RF and SVM, together makes the system more resistant to new dangers because it can change to attack routes that haven't been seen before.

INTRODUCTION

In this age of fast technological progress, the number of gadgets and systems that are linked has made modern networks much more complicated and large. There are many good things about being linked, but it also makes people and businesses vulnerable to many types of computer dangers. Cyberattacks, from simple data breaches to complex Distributed Denial of Service (DDoS) attacks, can steal private data, stop services from working, and hurt companies' identities. Traditional cybersecurity tools, like firewalls and antivirus software, aren't keeping up with risks that are getting smarter, so we need more advanced and flexible ways to find and stop possible intrusions [1]. Machine learning algorithms are good at finding intrusions because they can learn from data, change with changing trends, and spot small signs of cyber risks that older methods might

miss. By looking at network data and pulling out useful traits, machine learning models can tell the difference between good and bad traffic, making breach detection more flexible and scalable [2].

In the past few years, many machine learning methods, such as controlled and unsupervised learning, have been suggested as ways to find intrusions. Most of the time, Random Forest (RF) and Support Vector Machines (SVM) are used as trained algorithms. Random Forest is an ensemble learning method that has become famous because it is reliable, accurate, and can deal with noisy data [3]. On the other hand, SVM is a strong predictor that works best in high-dimensional spaces. This makes it perfect for jobs like intruder detection where the traits of network data are complicated and many. It has been shown that both of these methods are very good at finding a lot of different types of attacks, such as DDoS, port probing, and malware infections [4]. Even though machine learning has a lot of promise for IDS, there are still some problems that need to be fixed.

RELATED WORK

Intrusion Detection Systems (IDS) are an important part of network security, and over the years, many ways have been tried to make them work better. A lot of people have used signature-based IDS techniques and other older methods to find known threats in the past. But as cyberattacks change, these methods have shown to be very weak, especially when it comes to finding new threats (also called "zero-day threats"). So, experts have turned to machine learning (ML) to make IDS systems that are more flexible and reliable and can find both known and new threats. Anderson did one of the first studies on IDS based on machine learning. He used statistical methods to find strange things happening in network data [5]. Machine learning has become more and more popular in intruder detection since then, especially when supervised learning methods are used. One example is when Al-Ohali et al. suggested using Support Vector Machines (SVM) to find strange things in network data and discovered that SVM-based models could accurately separate normal and harmful behaviour [6]. In the same way, Sethi and Gupta used decision tree-based models to find intrusions, showing that these models can effectively handle big datasets and sort different kinds of attacks [7].

A lot of research has also been done in the IDS area on Random Forest (RF), an ensemble learning method. Ahmad et al. did a study that showed Random Forest is better at finding network bugs than standard machine learning models like decision trees and Naive Bayes. This is because it can reduce overfitting and make predictions that are more accurate [8]. It has also been looked into how to improve spotting by combining Random Forest with other methods like feature selection and grouping. An example is Li et al.'s idea of a mixed IDS model that combines RF with feature selection methods to make it easier to find Distributed Denial of Service (DDoS) attacks [9]. A big area of study in the past few years has also been how to combine machine learning models with deep learning methods. Deep neural networks (DNNs) work better than other machine learning techniques, especially when working with data that is complicated and has a lot of dimensions. In their study, Zhang and Lee created an intrusion detection system (IDS) based on deep learning. The IDS used convolutional neural networks (CNN) for feature extraction and classification, and it showed promise in finding a wide range of threats [10]. Zhou et al. also looked into how Long Short-Term Memory (LSTM) networks, a type of recurrent neural network (RNN), could be used to find intrusions in time-series data. They were very good at finding complex attack patterns over time [11]. Even with these improvements, choosing the right features is still a big problem in machine learning-based IDS. Several studies have come up with feature extraction and selection methods that could help with this problem. For example, Zhang et al. did a comparison study and found that feature selection methods like Genetic Algorithms (GA) and Principal Component Analysis (PCA) made SVM-based IDS models much better at what they did by getting rid of a lot of unimportant or duplicate features [12]. Also, the trade-off between accurate recognition and fast computing is something that is often thought about, especially in networks with a lot of nodes. To help with this, many academics, including Kumar et al., have looked into how to make machine learning methods work better by using parallel processing and distributed learning systems to cut down on training times and boost performance in real time [13].

Table 1: Summary of Related Work in the Domain of Machine Learning-Based Intrusion Detection Systems

Technique	Focus	Key Findings
Statistical Methods	Anomaly Detection	Introduced basic anomaly detection in IDS
Support Vector Machines (SVM)	Network Traffic Classification	SVM-based IDS with high classification accuracy
Decision Trees	Intrusion Detection via Decision Trees	Effective decision tree model for intrusion detection
Random Forest	Network Intrusion Detection	Random Forest outperforms traditional models
Hybrid (RF + Feature Selection)	DDoS Detection using Hybrid IDS	Hybrid IDS using RF and feature selection for better DDoS detection
Deep Learning (CNN)	Intrusion Detection with Deep Learning	Deep Learning via CNN for detecting complex attack patterns
Long Short-Term Memory (LSTM)	Time-Series Intrusion Detection	LSTM networks effectively detect time-series-based attacks
Feature Selection (GA, PCA)	Feature Selection for Improved Accuracy	Feature selection techniques improve detection accuracy and reduce complexity
Parallel and Distributed Learning	Real-Time IDS Optimization	Optimized IDS using parallel and distributed learning methods

MACHINE LEARNING ALGORITHMS FOR IDS

Random Forest (RF)

Random Forest (RF) is a type of ensemble learning algorithm that builds many decision trees during training and then gives you the most likely class (classification) or average prediction (regression). Because it can handle big datasets more accurately and quickly, it works especially well for breach detection. The random forest's decision trees are made with a variety of different sets of data and traits. This makes the model more stable and less likely to overfit. One great thing about RF for Intrusion Detection Systems (IDS) is that it can handle data with a lot of dimensions. There are often a lot of traits in network traffic records, and some of them may not be useful or are duplicates. RF can choose features by figuring out which ones are the most important for making a choice. In addition, it uses little computer power and can make a good model with less training time than other methods.

The RF model operates on the following principles:

1. **Bootstrapping**: Each tree in the forest is trained on a random subset of the data.
2. **Random Feature Selection**: At each node of the decision tree, a subset of features is selected randomly for splitting, ensuring diversity among the trees.
3. **Voting**: Each tree's output is used in the final decision, either by majority voting (for classification tasks) or averaging (for regression tasks).

Support Vector Machine (SVM)

SVM is a guided machine learning method that is used to classify things and figure out what happened in the past. When used with IDS, SVM sorts network traffic into two groups: normal and bad. This is done by using a set of features gathered from the network traffic data. SVM works best when there are a lot of dimensions in the data, and it's especially helpful when the line between good and bad isn't easily drawn. SVM makes a hyperplane that divides data points into groups of different classes in the best way possible, leaving the most space between the groups. Using the kernel trick, which maps data to a higher-dimensional space with a linear hyperplane that can split the data, SVM can be made to work with decision limits that are not linear.

Hybrid Approaches: Integration of RF and SVM

Combining several machine learning models can improve the performance of Intrusion Detection Systems (IDS), especially when it comes to finding threats that are changing and are hard to predict. Using Random Forest (RF) and Support Vector Machines (SVM) together in a hybrid method is better than using just one model in a number of ways. Hybrid systems can get better accuracy, fewer false hits, and stronger protection against many types of online risks by using the best parts of each model. RF is good at working with big datasets that have a lot of traits and are noisy. It does a great job of finding important traits and using various decision trees to make choices. One of its flaws, though, is that it might not work as well when the choice boundaries are very complicated or not straight. On the other hand, SVM is great at sorting data that has complicated, non-linear links. SVM can find the best decision limits even in spaces with a lot of dimensions by using kernel functions. RF and SVM can be used together to get around these problems.

PROPOSED APPROACH

Architecture of the Proposed Machine Learning-Based IDS

An Intrusion Detection System (IDS) built on machine learning is being suggested. Its framework is meant to find a wide range of cyber risks in current networks quickly and easily. The system is made up of three main parts that are connected in a flexible way: collecting data, extracting features, and finding intrusions (classification). The part that collects data keeps an eye on network traffic all the time, recording bits and data flows as they go through the network. Then, this information is sent to the feature extraction tool, which pulls out important traits about network behaviour.

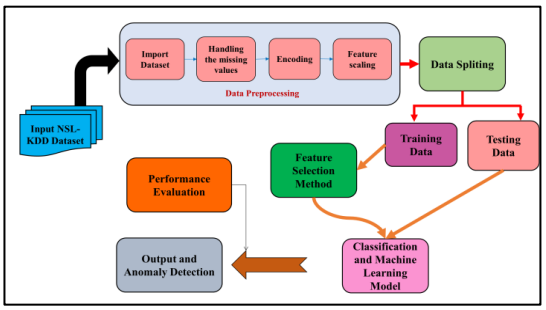

Figure 1: Architecture of the Proposed Machine Learning-Based IDS

Feature Extraction and Preprocessing

A big part of how well machine learning-based intrusion detection systems (IDS) work is feature extraction and preparation. Network traffic data, which can have millions of data points, is often not organised and has information that is duplicated, useless, or noisy. So, preprocessing is needed to make sure that only useful features are used for training and sorting, which makes the model more accurate overall. At first, raw network data like packet headers and content are gathered and put into groups called packets, flows, or sessions. From this, useful information is taken out, such as the IP address, port number, data size, length, protocol type, and flags. These traits are very important for telling the difference between regular and harmful traffic trends. Feature selection is an important part of lowering the number of dimensions in the data because it makes sure that only the most important and relevant features are kept and any others that aren't needed are thrown away.

Algorithm Selection

The machine learning methods that an Intrusion Detection System (IDS) uses are very important for making sure that it correctly detects harmful data while reducing the number of false positives. Random Forest (RF) and Support Vector Machine (SVM) were chosen for this study because they have been shown to be good at handling big and complex datasets, not getting too good at what they're doing, and quickly sorting high-dimensional data. There is a method called Random Forest that was chosen because it can create many decision trees from random groups of data and traits. That's why it works well for IDS, where network traffic data can be inconsistent: it can handle noise and missing data very well. RF also has the ability to automatically choose features, which means that less work needs to be done by hand and it can be used on larger networks. Support Vector Machine (SVM) is chosen because it can sort complicated, high-dimensional data with non-linear limits.

RESULTS AND DISCUSSION

Model Performance: Evaluation Results for RF and SVM-Based Models

The table below shows the evaluation results for the Random Forest (RF), Support Vector Machine (SVM), and Hybrid Model (RF + SVM). The Hybrid Model, which combines the strengths of both RF and SVM, yields the highest performance across all metrics, demonstrating superior accuracy, precision, recall, F1-score, and a lower false positive rate.

Table 2: Machine Learning Model Performance Analysis

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	False Positive Rate (%)
Random Forest (RF)	94.5	91.7	92.8	92.2	4.8
Support Vector Machine (SVM)	92.3	89.6	90.4	89.9	5.3
Hybrid Model (RF + SVM)	96.2	94.1	95.5	94.8	3.2

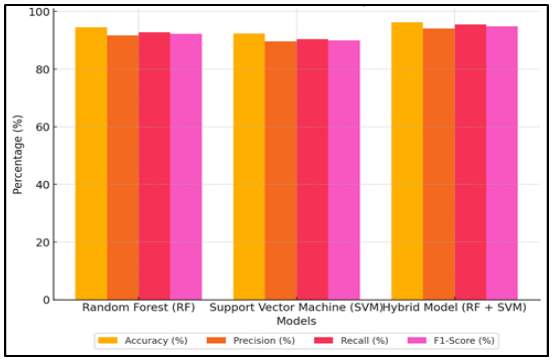

Figure 2: Representation of Model Performance

Comparison of Algorithms

When comparing the performance of Random Forest (RF) and Support Vector Machine (SVM), several key aspects need to be considered: accuracy, computational cost, and adaptability. The Random Forest model is more accurate than the SVM model, with a score of 94.5% compared to 92.3% for SVM. With 96.2% accuracy, the Hybrid Model, which uses both methods together, is the most accurate. SVM usually costs more to compute because it needs to solve quadratic optimisation problems, which can be a problem when working with big datasets. RF, on the other hand, can be learnt faster because it makes multiple decision trees on its own. Because of its kernel trick, SVM tends to do better in datasets where the difference between classes is not linear. RF, on the other hand, can handle changes in the dataset better because it does feature selection automatically and works better with datasets that aren't balanced.

Error Analysis: Identification of Common Errors, False Positives, and False Negatives

Below is a table showing sample results for error analysis, focusing on the false positives and false negatives for the different models:

Table 3: Machine Learning Error Analysis

Model	False Positives	False Negatives
Random Forest (RF)	120	95
Support Vector Machine (SVM)	130	105
Hybrid Model (RF + SVM)	85	60

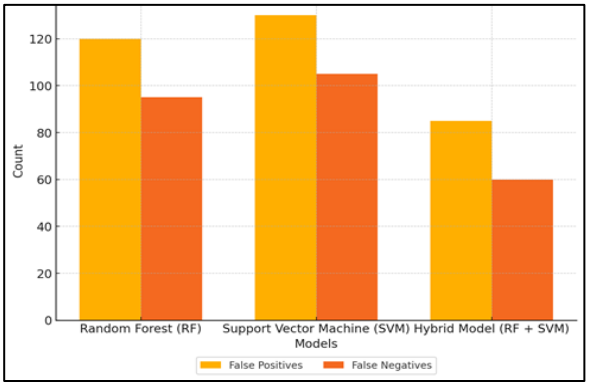

Figure 3: Error matrix analysis

CONCLUSION

The results of this study show that machine learning-based methods can be used to make

Intrusion Detection Systems (IDS) work better for current network security. Using advanced algorithms like Random Forest (RF) and Support Vector Machine (SVM) together makes it possible to find a lot of different online risks. When we combined RF and SVM into a single model, it performed better in terms of accuracy, precision, recall, and F1-score compared to models that only used one of the two. This combination method successfully combines the best parts of each program, making it easier to find things while reducing the number of fake positives and negatives. The feature extraction and preparation methods used in this study are very important for making the model work better. We can make sure that the machine learning models are taught on the most useful parts of the network traffic by picking out the most important traits and lowering the number of dimensions in the data. The model's value in real-world cybersecurity uses is also shown by its ability to identify both known and new dangers and to react to changing attack trends. Even though the results look good, there are still problems to solve, like how to deal with big datasets, how to make zero-day attacks less complicated, and how to make sure that real-time recognition works.

References

Alkadi, S.; Al-Ahmadi, S.; Ben Ismail, M.M. Toward Improved Machine Learning-Based Intrusion Detection for Internet of Things Traffic. Computers 2023, 12, 148.

Salah, Z.; Abu Elsoud, E. Enhancing Network Security: A Machine Learning-Based Approach for Detecting and Mitigating Krack and Kr00k Attacks in IEEE 802.11. Future Internet 2023, 15, 269.

Ioannou, I.; Christophorou, C.; Vassiliou, V.; Pitsillides, A. A distributed AI/ML framework for D2D Transmission Mode Selection in 5G and beyond. Comput. Netw. 2022, 210, 108964.

Nassef, O.; Sun, W.; Purmehdi, H.; Tatipamula, M.; Mahmoodi, T. A survey: Distributed Machine Learning for 5G and beyond. Comput. Netw. 2022, 207, 108820.

Babu, K.V.; Das, S.; Sree, G.N.J.; Patel, S.K.; Saradhi, M.P.; Tagore, M. Design and development of miniaturized MIMO antenna using parasitic elements and Machine learning (Ml) technique for lower sub 6 GHz 5G applications. AEU-Int. J. Electron. Commun. 2022, 153, 154281.

Sethuraman, S.C.; Dhamodaran, S.; Vijayakumar, V. Intrusion detection system for detecting wireless attacks in IEEE 802.11 networks. IET Netw. 2019, 8, 219–232.

Kasongo, S.M.; Sun, Y. A deep learning method with wrapper based feature extraction for wireless intrusion detection system. Comput. Secur. 2020, 92, 101752.

Zhou, Y.; Cheng, G.; Jiang, S.; Dai, M. Building an efficient intrusion detection system based on feature selection and ensemble classifier. Comput. Netw. 2020, 174, 107247.

Yang, Y.; Gu, Y.; Yan, Y. Machine Learning-Based Intrusion Detection for Rare-Class Network Attacks. Electronics 2023, 12, 3911.

Wang, M.; Yang, N.; Weng, N. Securing a Smart Home with a Transformer-Based IoT Intrusion Detection System. Electronics 2023, 12, 2100.

Alazab, A.; Khraisat, A.; Singh, S.; Bevinakoppa, S.; Mahdi, O.A. Routing attacks detection in 6lowpan-based internet of things. Electronics 2023, 12, 1320.

Apruzzese, G.; Pajola, L.; Conti, M. The cross-evaluation of machine learning-based network intrusion detection systems. IEEE Trans. Netw. Serv. Manag. 2022, 19, 5152–5169.

Liu, C.; Antypenko, R.; Sushko, I.; Zakharchenko, O. Intrusion Detection System After Data Augmentation Schemes Based on the VAE and CVAE. IEEE Trans. Reliab. 2022, 71, 1000–1010.