ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

Network Intrusion Detection Using Machine Learning

Lakshmeswari Chenngiri[1], Talupula Sivaparvathi[2], Kurapati Pujitha[3], Naga Jyothi[4], R.Venkatesh [5]

[1] Assistant Professor, Department of Computer Science Engineering, Chalapathi Institute of Engineering and Technology, Chalapathi Rd, Nagar, Lam, Guntur, Andhra Pradesh- 522034

[2,3,4,5] Students, Department of Computer Science Engineering, Chalapathi Institute of Engineering and Technology, Chalapathi Rd, Nagar, Lam, Guntur, Andhra Pradesh- 522034

Email id: lakshmi.chennagiri@gmail.com[1], talupulasivaparvathisp@gmail.com[2], pujithakurapati2003@gmail.com[3], nagajyotohi6301@gmail.com[4], venkateshvenky93924@gmail.com[5]

Abstract:

The exponential growth of digitalization and data volumes, the cybersecurity threat landscape has become increasingly complex, amplifying the need for robust intrusion detection systems (IDS). Traditional IDS approaches often struggle with static architectures, requiring costly and frequent retraining to keep up with evolving threats. This study introduces an incremental, majority-voting IDS system that leverages machine learning to adapt to continuous network traffic streams without the need for extensive retraining. By integrating multiple machine learning algorithms—K-Nearest Neighbors (KNN), Logistic Regression, Bernoulli Naive Bayes, and Decision Tree classifiers—the system employs a collective decision-making approach to enhance detection accuracy and minimize false alarms in real-time. Results indicate that this multi-algorithm IDS framework offers substantial improvements in adaptability, performance, and resilience against intrusions, especially within real-world, imbalanced data scenarios.

Keywords: Intrusion Detection System (IDS), Cybersecurity, Network security, K-Nearest Neighbors (KNN)

1. Introduction

In today's highly interconnected digital environment, ensuring the security of computer networks is more critical than ever. The proliferation of internet-enabled devices, cloud computing, and online services has drastically expanded the attack surface, making network infrastructures increasingly vulnerable to a wide range of cyber threats. Among these threats, unauthorized intrusions such as denial-of-service (DoS) attacks, data breaches, malware

infections, and phishing attempts pose significant risks to the confidentiality, integrity, and availability of sensitive information.

Traditional network security mechanisms, such as firewalls and signature-based intrusion detection systems (IDS), are often limited in their ability to detect novel or sophisticated attacks. These systems rely on predefined rules or known attack patterns, making them ineffective against zero-day exploits and polymorphic malware. Consequently, there is a growing need for intelligent and adaptive solutions capable of identifying both known and unknown intrusions in real time.

Machine learning (ML) has emerged as a powerful tool for addressing these challenges in the field of network security. By learning patterns from large volumes of network traffic data, ML-based intrusion detection systems can generalize from past experiences to detect anomalies or malicious behavior effectively. These systems can automatically update their detection models as new types of threats are encountered, making them more robust and scalable compared to traditional methods.

This study explores the application of machine learning techniques for network intrusion detection, focusing on feature selection, data preprocessing, classification algorithms, and performance evaluation. By leveraging supervised and unsupervised learning methods, the goal is to develop a system capable of accurately detecting intrusions with minimal false positives, even in complex and dynamic network environments.

2. Literature review

Cybersecurity, in the current era, has emerged as an international imperative, driven by the critical need to protect systems from unwanted, unauthorized, and unforeseen interference [1]. These interferences can range from data breaches and information theft to threats that undermine the integrity and functionality of systems. Safeguarding against such threats is paramount in ensuring the smooth operation of systems, protecting sensitive data, and preserving user trust [2]. Intrusion Detection Systems (IDS) have traditionally served as a cornerstone of perimeter security [3, 4].

These systems are crafted with the purpose of uncovering and responding to suspicious or malicious activities within a network or system. Nevertheless, the conventional signature-based intrusion detection methods, reliant on established attack patterns and signatures, have been found lacking in the face of ever-evolving and sophisticated cyber threats. These solutions are

engineered to identify and react to questionable or potentially harmful actions occurring within a network or system. Nevertheless, conventional intrusion detection methods, which hinge on established attack patterns and signatures, have demonstrated their inadequacy when confronted with ever-changing and increasingly complex cyber threats [5, 6].

To address the shortcomings of traditional IDS, the cybersecurity community has turned its attention to Machine Learning (ML) as a promising solution. ML-enabled IDS leverages behavior analysis to detect anomalies and threats, offering the potential for significantly higher accuracy and faster detection times [7, 8]. This paradigm shift in intrusion detection holds the promise of not only bolstering security but also reshaping the privacy landscape. This shift towards ML-enabled intrusion detection has sparked concerns regarding both privacy and the field of data science [9, 10]. ML algorithms, while effective at identifying threats, often require access to sensitive data. Balancing the need for security with privacy concerns is a challenge that demands innovative and ethical solutions

3. Methodology

Data preprocessing is a crucial part of any ML model. Models without preprocessing can create problems with invalid, overfitting, generating error models, providing low accuracy and much more. So, preprocessing is a very significant part of an ML model. To analyse our model, we have used some preprocessing techniques such as: handling the missing value by eradicating rows containing null, -inf and inf values, removing space from columns names to work with columns smoothly and dropping the duplicate rows by keeping the first one and delete the rest from the dataset, merge the similar classes with low instance from output columns and finally, reduce the dataset size by converting data types from int64 to int32 and foat64 to foat32 to train models with less dataset size but same dataset entries.

Table: The frequency distribution of attack categories of the CIC-IDS2018 dataset

Attack categories	Count	(%) Percentage
Benign	6,58,454	70.553
DDOS attack-HOIC	68,601	7.351
DDoS attacks-LOIC-HTTP	57,619	6.174
DoS attacks-Hulk	46,191	4.949
Bot	28,619	3.067
FTP-BruteForce	19,336	2.072
SSH-Bruteforce	18,759	2.01
Infilteration	16,193	1.735
DoS attacks-SlowHTTPTest	13,989	1.499
DoS attacks-GoldenEye	4151	0.445
DoS attacks-Slowloris	1099	0.118
DDOS attack-LOIC-UDP	173	0.019
Brute Force -Web	61	0.007
Brute Force -XSS	23	0.002
SQL Injection	9	0.001
Total	9,33,277	100

ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

Decision Tree (DT)

A non-parametric supervised ML technique called the DT is used to solve problems with regression and classification. The prediction of the value of the output of a dataset is generated gripping decision rules from dataset features. It's easy to comprehend and interpret and it can be visualized. It can handle multi-output problems [54]. It is widely used in IDS. Decision node, having multiple branches and confirmed to make the decision and Leaf nodes, not contain any branches and the output of those decisions are the two nodes in DT. The starting decision node is called the root node. To build a decision tree, Attribute Selection Measure (ASM) is performed on information gain and the Gini index to select the feature [55]. The change of entropy based on a feature after splitting is called IG. Based on the value of IG, we have separated the node and constructed the decision tree based on the value of IG. Te measure of purity or impurity creating a DT is called GI. To create binary splits, GI is used

Environment setup

The experiments are conducted in a robust computing environment, utilizing a high performance 2X-large virtual machine instance. This instance boasts 8 cores, allowing for efficient concurrent task handling and enhanced multi-threading capabilities. With 64 GB of RAM, the system is well-equipped to accommodate memory-intensive applications, and it offers a generous 40 GB of disk space for data storage. The experiments are seamlessly executed using the Jupiter notebook through Anaconda Navigator. To support our performance evaluation, we leverage the Python programming language and a suite of indispensable libraries, including TensorFlow, Kera's, Pandas, Scikit-learn, NumPy, Seaborn, Matplotlib, Imbalanced-learn etc.

Performance evaluation metrics

Several measures, such as accuracy, precision, recall, F1-score, ROC curve, and confusion matrix, are used to evaluate the performance of our proposed model. The following defines these performance matrices:

Confusion matrix

The Confusion Matrix is a valuable tool for evaluating ML classification performance. It is a tabular representation containing four combinations of predicted and actual values: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) illustrates a

confusion matrix where TP represents correctly anticipated positive values, TN indicates accurately projected negative values, FP corresponds to incorrectly

Results

The performance results for binary and multilabel classification on the CIC-IDS2017 dataset are presented in Fig. the inclusion of both scenarios allows for a comprehensive assessment of model performance on the CICIDS2017 dataset in the bar chart, it's evident that our proposed model exhibits a substantial increase in accuracy for binary and multilabel classification. Interestingly, the rate of accuracy improvement is notably higher in multilabel classification when compared to binary classification. In the binary classification, the accuracy rates on the proposed model are as follows: 99.91% for DT, 99.94% for RF, 99.95% for ET, and 99.65% for XGB. Te accuracy rates for multilabel classification are 99.91% for DT, 99.94% for RF, 99.95% for ET, and 99.65% for XGB on the proposed model

Table: Performance metrics for binary classification for CIC-IDS2017 dataset

ML	Accuracy		Precision		Recall		F1-score	
	All feature	Proposal	All feature	Proposal	All feature	Proposal	All feature	Proposal
DT	99.87	99.91	99.78	99.91	99.78	99.91	99.78	99.91
RF	99.9	99.94	99.82	99.94	99.82	99.94	99.82	99.94
ET	99.83	99.95	99.73	99.95	99.7	99.95	99.7	99.95
XGB	99.92	99.65	99.83	99.65	99.86	99.65	99.86	99.65

Table: Performance analysis of multilabel classification for CIC-IDS2017 dataset

ML	Accuracy		Precision		Recall		F1-score	
	All feature	Proposal	All feature	Proposal	All feature	Proposal	All feature	Proposal
DT	99.85	99.99	98.69	99.99	94.69	99.99	94.69	99.99
RF	99.89	99.99	98.98	99.99	94.17	99.99	94.17	99.99
ET	99.83	99.99	98.57	99.99	94.14	99.99	94.14	99.99
XGB	99.92	99.94	99.3	99.94	94.47	99.94	94.47	99.94

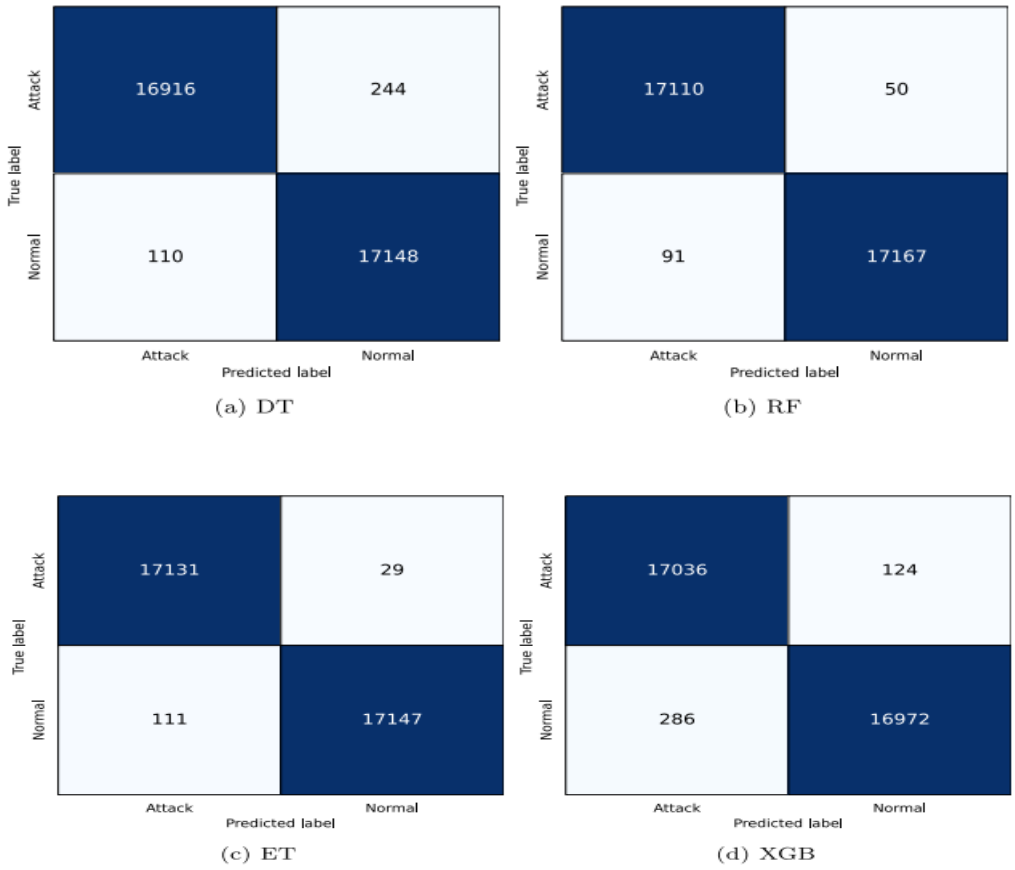

Fig: Binary confusion matrix for UNSW-NB15 dataset

Conclusion

In conclusion, our research has introduced a novel approach to network intrusion detection by combining various techniques to address the challenges of imbalanced data, feature embedding, and dimension reduction. Our model leverages the Random Oversampling (RO) method to tackle data imbalance, utilizes feature embedding through K means and GM clustering results, and employs Principal Component Analysis (PCA) for dimension reduction The limitation of our research is that we did not employ deep learning models along with optimization techniques. While our current approach has demonstrated remarkable results, there remains untapped potential for further improving the performance of intrusion detection systems. In the future, we envision expanding our work to incorporate deep learning models, which have shown great promise in various felds, including intrusion detection. DL algorithms, such as DNN, RNN or Hybrid are capable of capturing intricate patterns and representations in comp

References

1. Mueller S. Facing the 2020 pandemic: what does cyberbiosecurity want us to know to safeguard the future? Biosaf Health. 2021;3(1):11–21.
2. Marwala T. Cybersecurity in politics. In: Artifcial intelligence, game theory and mechanism design in politics. Springer; 2023. p 135–155.
3. George AS, George AH, Baskar T. Digitally immune systems: building robust defences in the age of cyber threats. Partners Univ Int Innov J. 2023;1(4):155–72.
4. Nguyen H, Lim Y, Seo M, et al. Strengthening information security through zero trust architecture: a case study in South Korea. In: International conference on intelligent systems and data science, Springer;2023 pp 63–77.
5. Khan A, Rehman M, Rutvij H, Jhaveri R, Raut T, Saba SA. Deep learning for intrusion detection and security of Internet of things (IoT): current analysis, challenges, and possible solutions. Security and Communication Networks. 2022.
6. Talukder MA, Hasan KF, Islam MM, et al. A dependable hybrid machine learning model for network intrusion detection. J Inf Secur Appl. 2023;72(103):405
7. Schmitt M. Securing the digital world: protecting smart infrastructures and digital industries with artifcial intelligence (ai)-enabled malware and intrusion detection. J Ind Inf Integr. 2023;36(100):520.
8. Preuveneers D, Joosen W. Sharing machine learning models as indicators of compromise for cyber threat intelligence. J Cybersecur Priv. 2021;1(1):140–63.
9. Singh P, Singh P. Artifcial intelligence: the backbone of national security in 21st century. Tuijin Jishu/J Propul Technol. 2023;44(4):2022–38.
10. Mohammadi S, Mirvaziri H, Ghazizadeh-Ahsaee M, et al. Cyber intrusion detection by combined feature selection algorithm. J Inf Secur Appl. 2019;44:80–8.