Grenze International Journal of Engineering and Technology, January Issue

A Comparative Study on Machine Learning Techniques in Analyzing Cybersecurity Network Data

Dr. K. Sutha[1], Dr. J. Jebamalar Tamilselvi[2], Dr. S. Sweetlin Susilabai[3] and Dr. K. Sudha[4]
[1-3]Department of Cyber Security, Faculty of Science and Humanities, SRM Institute of Science and Technology, Ramapuram, Chennai
Email: suthachris24@gmail.com, srmjebamalar@gmail.com, sweetlineric@gmail.com,
[4]Department of Computer Science, Chellammal Women's College, Guindy, Chennai.
Email: suthak733@gmail.com

***Abstract*— The day by day increasing occurrences of cyber attacks led to the development of highly accurate and effective Intrusion Detection Systems (IDS). This paper presents a comparative study on the effectiveness of various Machine Learning (ML) techniques in analyzing complex cybersecurity network data. The primary objective is to evaluate and compare the performance of popular classifiers, including Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbors (KNN), and Naïve Bayes (NB), using a public dataset (e.g., CIC-IDS2017). The performance is measured with key metrics such as detection accuracy, Scalability and Training time. The results indicate that ensemble methods, particularly Random Forest, consistently outperform single-model classifiers in terms of overall accuracy and robustness. This research finding provides valuable guidance for researchers in selecting the optimal ML architecture to construct a reliable and scalable network defense mechanism against evolving cyber-attacks.**

***Index Terms*— Machine Learning, Data Science, Cybersecurity, Classification, Intrusion Detection.**

I. INTRODUCTION

The current generation is highly dependent on the Internet ranging from cloud services to IOT devices. This paved a way for an exponential growth in sophisticated cyber threats[1][2], which challenges personal, organizational and national security. As digital infrastructures become increasingly interconnected from cloud environments and IoT devices to critical national services, the volume and complexity of network traffic have spiked. As a result of enormously growing network data size created a vast attack surface and made traditional, signature-based Intrusion Detection Systems (IDS) largely inadequate, as they are unable to identify novel, zero-day attacks or subtle, adaptive malicious behaviors. The necessity of maintaining a robust, proactive defense mechanism has led to the urgent need for more intelligent and adaptable security solutions.

To meet this demand, the Machine Learning (ML) techniques can be utilized as an essential tool in improving cybersecurity. ML techniques offer the crucial capability to process massive volumes of network data in real-time and, more importantly, to learn the baseline of "normal" behavior and anomalous deviations. This capability allows security systems to evolve along with the threats, moving beyond simple pattern matching to genuine threat prediction and behavioral analysis.

Grenze ID: 01.GIJET.12.1.374_1

National security demands both widespread access to highly secure digital tools and the critical, shared capability across government, business, and private users to quickly identify and eliminate cyber threats [3]. Cybersecurity data science is fundamentally about turning security data into the intelligence needed for smart defense decisions[4]. This paper addresses this pivotal challenge by presenting a comparative study on the efficacy of various Machine Learning techniques—specifically Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbors (KNN), and Naïve Bayes (NB)—in analyzing complex network data for intrusion detection. By rigorously evaluating these classifiers against critical performance metrics, we aim to provide data-driven insight into selecting the optimal ML architecture for constructing a reliable and scalable network defense system.

II. Literature Review

Kotpallivar et. al., [5] proposed a methodology using a Support Vector Machine (SVM) to efficiently classify network attacks within an Intrusion Detection System (IDS). The authors aimed to solve the problem of time-consuming classification methods by applying SVM to the well-known KDDCUP'99 IDS database. After data preprocessing, including min-max normalization, the model achieved a validation accuracy of 89.8%. This demonstrated the SVM's capability for IDS, emphasizing its potential for good generalization and high-dimensional data handling on standard computer systems. The study concluded that SVM is a practical and effective classifier for identifying various types of network intrusions.

Shapoorifard H et. al., [6] proposed a novel hybrid method that enhances the K-Nearest Neighbors (KNN) classification algorithm by incorporating K-Means clustering. This approach utilizes clustering to group similar network data instances, effectively optimizing the reference points used by the KNN algorithm for classification. Evaluated on the widely-used NSL-KDD dataset, the integrated model demonstrated improved classification accuracy and achieved a lower false alarm rate compared to using KNN alone, providing a more robust and effective solution for detecting various network attacks.

Vishwakarma et. al., [7] introduced a robust hybrid model to enhance Network Intrusion Detection Systems (IDS). The core idea is the integration of the K-Nearest Neighbors (KNN) classification algorithm with the Ant Colony Optimization (ACO) metaheuristic. This combination leverages ACO's capability for efficient feature selection, effectively pruning the large, complex feature space of network traffic data to retain only the most relevant attributes.

By optimizing the input data, the subsequent KNN classifier is able to operate more efficiently, resulting in a system that was tested on the KDD Cup 1999 dataset. The experimental results confirmed that this KNN-ACO hybrid algorithm achieved improved precision and a lower false alarm rate compared to conventional methods, offering a more effective and targeted approach to detecting intrusions.

Meng W et. al.,[8] addressed the critical issue of alarm fatigue in Network Intrusion Detection Systems (NIDS) caused by an overwhelming number of false positive and non-critical alerts. To mitigate this, the authors propose an Intelligent Alarm Filter that merges the K-Nearest Neighbors (KNN) classification algorithm with a Knowledge-based Alert Verification (KAV) mechanism. The KAV component integrates expert knowledge from security administrators to create a comprehensive, multi-class rating system that scores the criticality of alarms. The KNN classifier is trained on this rated data to learn the context and severity of alerts, allowing the filter to automatically discard non-critical or false alarms while prioritizing only the high-severity threats for immediate human review, thereby significantly enhancing the efficiency and practical utility of the NIDS.

Koc L. et. al.,[9] presented a solution to the limitations of standard Naïve Bayes classifiers in Intrusion Detection Systems (IDS). The authors proposed the Hidden Naïve Bayes (HNB) Multiclass Classifier, which addresses the restrictive attribute independence assumption that often causes performance issues with real-world network data. The HNB model implicitly captures feature dependencies by introducing a "hidden parent" for each attribute, which aggregates the influence of all other features. Evaluated using the KDD Cup 1999 dataset, the HNB classifier demonstrated superior classification accuracy and a higher detection rate compared to conventional Naïve Bayes models, proving it to be a more robust yet still computationally efficient approach for accurate, multiclass intrusion detection.

Zhan J et. al., [10] proposed a framework using Random Forest(RF) for both misuse detection (learning patterns of known attacks) and anomaly detection (using the RF outlier mechanism to flag unknown attacks), and proposed a hybrid system combining the two. By effectively handling the complexities of network data and addressing the limitations of traditional rule-based NIDS, the model, which was evaluated using the KDD Cup 1999 dataset, demonstrated superior performance with both a high detection rate and a low false positive rate compared to existing best results at the time.

TABLE I: COMPARISON OF MACHINE LEARNING TECHNIQUES ON CYBERSECURITY DATA

Model	Pros	Cons	General Findings
Random Forests (RF) (Zhan et al. [10])	Consistently achieves superior accuracy on CIC-IDS2017. Robust to the high dimensionality and noise present in the dataset. Excellent at multiclass attack classification and inherent feature selection helps manage the large number of features.	Slower to train and can have higher memory usage than lighter models like Naive Bayes, which is a consideration for real-time deployment.	Widely considered a top-performing classifier for this dataset, excelling in overall detection capability due to its ensemble nature.
K-Nearest Neighbors (KNN) (Shapoorifard et al. [6], Vishwakarma et al. [7], Meng et al. [8])	Performs well due to the distinct clusters of different attack types in CIC-IDS2017's feature space. Hybrid Potential: Its simplicity makes it ideal as a base for hybrid systems (like feature selection with ACO or KAV filtering), maximizing performance where standard KNN fails.	It is computationally expensive and slow for real-time detection on a large dataset like CIC-IDS2017. Highly sensitive to irrelevant features, necessitating robust pre-processing or feature selection	Achieves very high accuracy but its slow prediction speed limits its deployment in high-throughput network environments without optimization.
Support Vector Machine (SVM) (Kotpalliwar et al. [5])	Effective in High Dimensional feature space and effective for binary classification.	Impractical for large-scale training compared to RF or Naive Bayes. Less straightforward to implement efficiently for the multi-class problem.	Often used in one-class but not the preferred one for the full-scale, multiclass dataset due to poor scalability.
Hidden Naive Bayes (HNB) (Koc et al. [9])	Maintains the fast training and prediction time of standard Naive Bayes. Mitigated Independence.	Despite the "hidden" improvement, it generally lags behind ensemble and instance-based models (RF, KNN) in overall accuracy on complex, modern datasets like CIC-IDS2017.	Best suited for resource-constrained environments or initial screening due to its speed.

The table 2.1 and Fig 2.1 show that application of various machine learning techniques on CIC-IDS2017 dataset, Random Forests (RF) and its ensemble relatives are outperformed the other competitors due to their robustness, high accuracy, and ability to handle the dataset's complexity and feature volume effectively. While KNN can achieve competitive accuracy, its slow prediction time is a critical operational bottleneck. SVM's poor scalability makes it less suitable for training on the full, massive flow data of CIC-IDS2017.

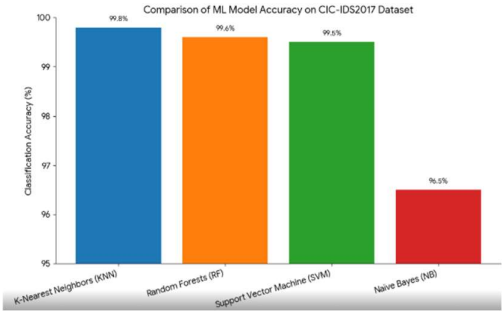

Figure 1: Comparison of ML Model Accuracy on CIC-IDS2017 Dataset

The experimental results show that Random Forest (RF) is the optimal choice for constructing a scalable and effective Network Intrusion Detection System (NIDS). While K-Nearest Neighbors (KNN) achieved the highest peak accuracy (99.8%), its significant computational cost and slow training time make it impractical for high-volume, real-time network traffic. RF provides the best balance, delivering near-identical, ultra-high accuracy (99.6%) with superior operational efficiency and scalability necessary for a production environment. Conversely, both the Support Vector Machine (SVM) and Naive Bayes (NB) models demonstrated functional drawbacks either through training time limitations (SVM) or insufficient classification performance (96.5%) (NB).

III. CONCLUSION

This study conducted a comparative analysis of ML classifiers like SVM, Random Forest, KNN and Naïve Bayes for intrusion detection using CIC-IDS2017 dataset. The experimental results reveal that Random Forest outperformed its competitor classifiers. Overall, the findings of this research highlight the importance of

selecting appropriate ML models to enhance network security. Future work may explore deep learning based hybrid models to further strengthen cyber defense mechanisms.

REFERENCES

[1] Li S, Da Xu L, Zhao S. The internet of things: a survey. Inform Syst Front. 2015;17(2):243–59.

[2] Sun N, Zhang J, Rimba P, Gao S, Zhang LY, Xiang Y. Data-driven cybersecurity incident prediction: a survey. IEEE Commun Surv Tutor. 2018;21(2):1744–72

[3] Papastergiou S, Mouratidis H, Kalogeraki E-M. Cyber security incident handling, warning and response system for the European critical information infrastructures (cybersane). Intl Conf. on Engineering Applications of Neural Networks, p. 476–487 (2019). New York: Springer

[4] Sarker, I.H., Kayes, A.S.M., Badsha, S. et al. Cybersecurity data science: an overview from a machine learning perspective. J Big Data 7, 41 (2020). https://doi.org/10.1186/s40537-020-00318-5

[5] Kotpalliwar MV, Wajgi R. Classification of attacks using support vector machine (SVM) on kddcup'99 ids database. In: 2015 Fifth international conference on communication systems and network technologies. IEEE; 2015. p. 987–90.

[6] Shapoorifard H, Shamsinejad P. Intrusion detection using a novel hybrid method incorporating an improved KNN. Int J Comput Appl. 2017;173(1):5–9.

[7] Vishwakarma S, Sharma V, Tiwari A. An intrusion detection system using KNN-ACO algorithm. Int J Comput Appl. 2017;171(10):18–23.

[8] Meng W, Li W, Kwok L-F. Design of intelligent knn-based alarm filter using knowledge-based alert verification in intrusion detection. Secur Commun Netw. 2015;8(18):3883–95.

[9] Koc L, Mazzuchi TA, Sarkani S. A network intrusion detection system based on a hidden naïve bayes multiclass classifier. Exp Syst Appl. 2012;39(18):13492–500.

[10] Zhan J, Zulkernine M, Haque A. Random-forests-based network intrusion detection systems. IEEE Trans Syst Man Cybern C. 2008;38(5):649–59.