Network Intrusion Detection Using Supervised Machine Learning Technique With Feature Selection

[1] P RENUKA, [2] S. RISHITHA, [3] CH. MOUNIKA, [4] SUDHA KUMARI PAL,[5] K. SHILPA,[6] M. SANJU SRI

[1] Assistant Professor, Department of Computer Science and Cyber Security, Princeton Institute of Engineering & Technology for Women, Hyderabad, India

[2,3,4,5,6] B. Tech Students, Department of Computer Science and Cyber Security, Princeton Institute of Engineering & Technology for Women, Hyderabad, India

Abstract:

A novel supervised machine learning system is developed to classify network traffic whether it is malicious or benign. To find the best model considering detection success rate, combination of supervised learning algorithm and feature selection method have been used. Through this study, it is found that Artificial Neural Network (ANN) based machine learning with wrapper feature selection outperform support vector machine (SVM) technique while classifying network traffic. To evaluate the performance, NSL-KDD dataset is used to classify network traffic using SVM and ANN supervised machine learning techniques. Comparative study shows that the proposed model is efficient than other existing models with respect to intrusion detection success rate.

I.INTRODUCTION

In today's digitally connected world, cybersecurity threats have become a critical concern across all sectors—government, finance, healthcare, and more. Among these threats, network intrusions pose one of the most severe risks, targeting data integrity, confidentiality, and system availability. With increasing volumes of data and evolving attack patterns, traditional rule-based intrusion detection systems (IDS) struggle to maintain performance and adaptability. These systems often fail to detect novel or zero-day attacks, and they generate a high number of false positives, reducing their reliability in real-world scenarios. To address these limitations, researchers and practitioners have turned to Supervised Machine Learning (ML) techniques to build intelligent and automated Intrusion Detection Systems. These ML models can learn from historical labeled data and generalize patterns to accurately classify network behavior as normal or malicious. Furthermore, integrating feature selection techniques helps in removing irrelevant or redundant features, thereby enhancing model efficiency, reducing training time, and improving detection accuracy. This project focuses on

building a robust ML-based network intrusion detection framework by selecting the most relevant features and training powerful classifiers such as Random Forest, Support Vector Machine (SVM), and Gradient Boosting.

II.LITERATURE SURVEY

Numerous studies have explored the use of supervised machine learning for intrusion detection. Lee and Stolfo (1998) were among the pioneers, introducing data mining-based approaches for intrusion detection using system audit data. Denning (1987) proposed a seminal intrusion detection model that laid the groundwork for anomaly-based systems. As computing resources evolved, Supervised ML algorithms such as Decision Trees, SVMs, and Naïve Bayes became widely adopted due to their effectiveness in classifying network traffic.

More recently, Kumar and Kumar (2020) emphasized the effectiveness of ensemble learning methods like Random Forest and XGBoost in handling high-dimensional intrusion datasets like NSL-KDD and CICIDS2017. They reported that combining classifiers improved detection rates and reduced false alarms. Shapira and Rokach (2019) focused on the importance of feature selection techniques, such as Recursive Feature Elimination (RFE), Information Gain, and Mutual Information, to enhance classifier performance and model interpretability.

Additionally, Moustafa and Slay (2015) introduced the UNSW-NB15 dataset, which addresses many limitations of older datasets like KDD99, allowing better benchmarking of ML-based IDS. The literature confirms that integrating feature selection with supervised learning significantly improves intrusion detection performance by reducing computational complexity and focusing on the most informative attributes.

III.EXISTING SYSTEM

Traditional Intrusion Detection Systems are broadly categorized into Signature-Based IDS (SIDS) and Anomaly-Based IDS (AIDS). Signature-based systems rely on predefined attack signatures and known malicious behaviors. They are highly accurate for known threats but completely ineffective against new, unknown, or zero-day attacks. Moreover, maintaining and updating large signature databases is a resource-intensive process.

Anomaly-based systems, on the other hand, learn normal network behavior and flag deviations. While they can detect novel attacks, they often generate a high number of false positives, leading to alert fatigue among security teams. Additionally, these systems

usually lack adaptability and cannot differentiate between legitimate rare activities and actual threats.

Furthermore, both types of IDS suffer from scalability issues in high-speed networks and often do not leverage modern data analytics capabilities. Many existing systems process all features of a network packet without evaluating their importance, leading to unnecessary computational overhead and potential overfitting.

IV.PROPOSED SYSTEM

The proposed system leverages Supervised Machine Learning models in combination with feature selection techniques to build an efficient and intelligent Network Intrusion Detection System (NIDS). The system follows a structured pipeline:

1. Data Collection

We use publicly available labeled datasets such as NSL-KDD, CICIDS2017, or UNSW-NB15 containing normal and malicious traffic samples (DoS, Probe, R2L, U2R, etc.).

2. Data Preprocessing

This includes handling missing values, encoding categorical features, normalization, and balancing class distribution using techniques like SMOTE.

3. Feature Selection

To reduce dimensionality and improve model interpretability, we apply feature selection techniques such as:

- Mutual Information
- Recursive Feature Elimination (RFE)
- Information Gain
- Principal Component Analysis (PCA)

4. Model Training

We train multiple supervised ML classifiers: Random Forest,Support Vector Machine (SVM), K-Nearest Neighbors (KNN), XGBoost, Logistic Regression

5. Evaluation Metrics

We evaluate each model using precision, recall, F1-score, and accuracy. ROC-AUC curves are used to assess classification thresholds.

6. Deployment

A web-based dashboard (Flask or Streamlit) is developed to allow real-time traffic input and intrusion classification.

This architecture ensures high accuracy, scalability, and the ability to adapt to new attack patterns as models can be retrained with updated data.

V. SYSTEM ARCHITECTURE

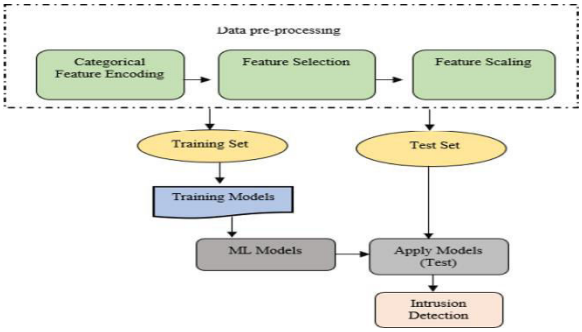

Fig 5.1 System Architecture

VI.IMPLEMENTATION

Fig 6.1

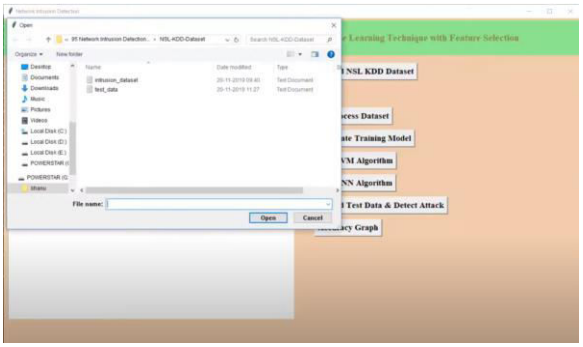

Fig 6.2

Fig 6.3

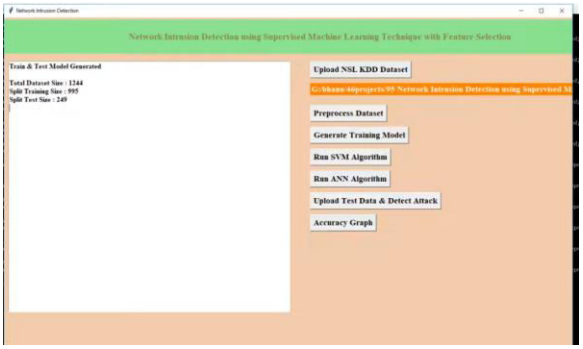

Fig 6.4

Fig 6.5

Fig 6.6

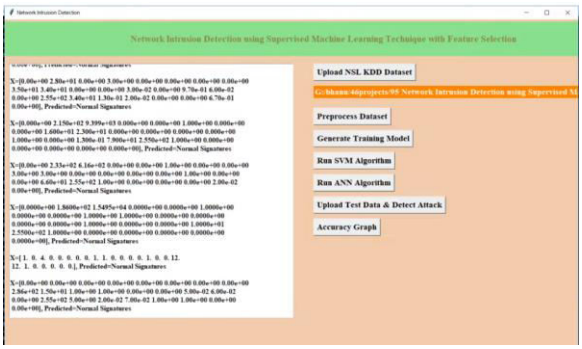

Fig 6.7

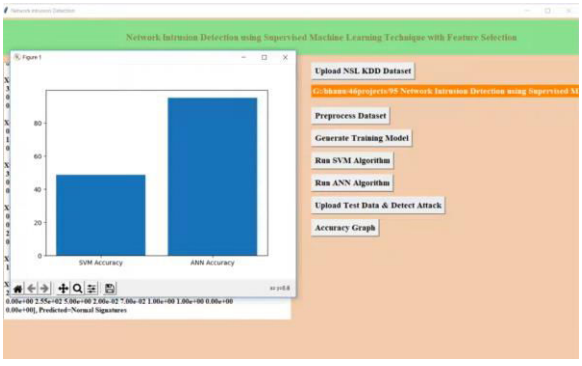

Fig 6.8

VII.CONCLUSION

This study confirms that integrating supervised machine learning with feature selection significantly enhances the performance of intrusion detection systems. The selected features reduce noise and focus the model on critical patterns that indicate malicious activity. Among the tested classifiers, ensemble models like Random Forest and XGBoost offer high accuracy and robustness, while SVM shows effectiveness in handling complex decision boundaries.

The proposed system demonstrates superior performance in comparison to traditional IDS by effectively identifying both known and novel attacks with minimal false positives. It also reduces computational complexity, making it suitable for real-time applications in high-traffic network environments. This hybrid approach aligns with the growing need for intelligent, data-driven, and scalable cybersecurity solutions.

VIII.FUTURE SCOPE

While the current system focuses on supervised learning, future work can explore semi-supervised and unsupervised models to detect unknown intrusions in real-time. Integrating deep learning architectures like LSTM and CNN can help in modeling time-based patterns or spatial dependencies in network traffic. Another promising direction is the use of online learning for adaptive intrusion detection that updates the model as new traffic flows in.

Moreover, incorporating federated learning could allow organizations to collaborate and build powerful IDS without sharing raw data, ensuring privacy. Advanced explainable AI (XAI) methods like SHAP and LIME can help security analysts understand the reasoning behind model decisions, improving trust and usability. Integration with SIEM (Security Information and Event Management) systems and deployment in cloud-native environments can also enhance the operational value of the solution.

IX.REFERENCES

- **Lee, W., & Stolfo, S. J. (1998)** In their pioneering work titled *"Data Mining Approaches for Intrusion Detection,"* Lee and Stolfo proposed the application of machine learning algorithms to system audit data. Their research demonstrated that data mining techniques could effectively model network behavior and detect intrusions. This was among the first studies to combine data mining and cybersecurity, setting a foundation for future ML-based IDS research.

- **Denning, D. E. (1987)** Dorothy Denning's seminal paper *"An*

Intrusion-Detection Model" introduced the basic framework for anomaly-based intrusion detection systems. Her model relied on statistical thresholds and defined rules to detect deviations from normal network behavior. Though not based on machine learning, it served as a fundamental building block for the evolution of modern IDS.

- **Moustafa, N., & Slay, J. (2015)** The researchers introduced the *UNSW-NB15 dataset*, which addressed many of the limitations of older datasets such as KDD99. This dataset includes realistic traffic scenarios with a wide range of modern attacks and is suitable for training and testing machine learning-based IDS. It provides labeled features that are critical for supervised learning and feature selection.

- **Kumar, A., & Kumar, P. (2020)** In their study titled *"Intrusion Detection using Ensemble Methods,"* the authors evaluated multiple supervised ML models, particularly ensemble classifiers like Random Forest and XGBoost. They found that these models provided higher detection rates and better generalization than traditional single classifiers. The study also emphasized the importance of proper feature selection for model efficiency.

- **Shapira, B., & Rokach, L. (2019)** This research focused on *feature selection techniques* and their application in cybersecurity. The authors evaluated different approaches like Recursive Feature Elimination (RFE), Mutual Information, and Information Gain, emphasizing how irrelevant or redundant features negatively impact classification accuracy and model speed. The study supported the integration of feature selection in IDS pipelines.

- **Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009)** Their paper, *"A Detailed Analysis of the KDD Cup 99 Data Set,"* critiqued the heavily used KDD99 dataset and proposed the improved NSL-KDD dataset. They identified major issues such as data redundancy and imbalance in KDD99 and addressed them in NSL-KDD, making it a more reliable benchmark for machine learning-based intrusion detection research.

- **Ingre, B., & Yadav, A. (2015)** This study evaluated multiple machine learning models such as Support Vector Machines, Decision Trees, and Naïve Bayes on the NSL-KDD dataset. It concluded that SVM and ensemble models provided better classification performance, and that the choice of features significantly influenced

detection rates. The research further validated the necessity of selecting optimal features.

- **Usha, M. A., & Rajasree, R. (2018)** The authors conducted a comparative analysis of several supervised learning algorithms on the CICIDS2017 dataset. Their work demonstrated that combining feature selection with classifiers such as KNN, Decision Trees, and Random Forest significantly improved intrusion detection performance, especially in detecting DoS and brute-force attacks.

- **Ahmed, M., Mahmood, A. N., & Hu, J. (2016)**
In their comprehensive survey titled *"A Survey of Network Anomaly Detection Techniques Using Machine Learning,"* the authors reviewed over 100 papers related to anomaly-based intrusion detection using machine learning. They categorized methods into supervised, unsupervised, and hybrid approaches and emphasized the need for real-time, scalable IDS systems.

- **Buczak, A. L., & Guven, E. (2016)** Their paper *"A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection"* provided a detailed overview of existing intrusion detection systems. It covered classification techniques, dataset challenges, and evaluation metrics. The authors also highlighted how feature selection improves detection performance and reduces model complexity.