Intrusion Detection System

Ch.Meghana[1], K.Rahul[2], N.Sai Venkat[3], Ch.Lohith[4], and T.Rahul[5]
[1]*Assistant Professor, Hyderabad institute of technology and management, Medchal, Telangana*
[2,3,4,5] *UG student, Hyderabad institute of technology and management, Medchal, Telangana*

Abstract-Intrusion Detection Systems are considered key elements in maintaining network security that scrutinize traffic to identify possible threats. This work envisions the construction of a machine-learning based IDS. The NSL-KDD dataset, one of the most common benchmarks for network intrusion detection, is used. Two algorithms: Isolation Forest for anomaly detection, and Random Forest for attack classification, were implemented along with a thorough evaluation of the metrics of accuracy, precision, and recall. Results The findings reveal the potential ability of machine learning approaches to improve the detection accuracy of intrusions with minimal false positives. The system, then, holds significant promise for real-time application within cybersecurity infrastructures.

Index Terms—Intrusion Detection System, Machine Learning, Anomaly Detection, NSL-KDD, Cybersecurity, Random Forest, Isolation Forest

I. INTRODUCTION

As the sophistication of cyber-attacks continues to evolve, the need for more adaptive and intelligent security systems becomes paramount. Traditional signature-based IDS systems are increasingly inadequate in detecting new or unknown attack patterns, as they depend on previously identified attack signatures. This limitation is particularly evident in the context of zero-day attacks, where attackers exploit vulnerabilities that are not yet known to security professionals. Machine learning (ML)-based IDS systems, however, offer the potential to bridge this gap by continuously learning from network traffic data and improving their detection capabilities over time. These systems do not rely solely on predefined signatures but rather analyze features within the data to identify deviations from normal behavior, which can signify new or emerging threats.

The use of machine learning models such as Isolation Forest and Random Forest in intrusion detection represents an innovative approach to cybersecurity. Isolation Forest is particularly effective for anomaly detection as it isolates outliers in the data by constructing decision trees, making it highly suited for identifying rare or novel attacks that deviate from normal network behavior. On the other hand, Random Forest, a powerful ensemble learning method, excels in classifying various types of attacks by aggregating predictions from multiple decision trees, thus increasing the accuracy and reliability of the system. By combining both models, this approach leverages the strengths of anomaly detection and attack classification, offering a robust IDS solution capable of identifying both known and unknown threats. The comparison of these models highlights their individual strengths and limitations, providing valuable insights for designing more efficient and accurate real-time intrusion detection systems.

II. LITERATURE SURVEY

Intrusion Detection System(IDS) play an important role in protecting a network from cyber attacks; however, the traditional systems based on signature are not effective while the cyber attacks are evolving as fast as a tornado. The advent of ML has changed the landscape of IDS altogether, now capable of providing adaptive, scalable, and precise detection mechanisms to counter the weaknesses of traditional systems.

Supervised Learning Techniques

One such supervised learning techniques used by IDS are Decision Trees, SVM, or Random Forest to classify network traffic. In an ensemble method called Random Forest, predictions made by many decision trees are combined in order to improve accuracy with much less chance of [1]. These models are quite efficient for detecting known patterns of attacks, assuming there is an appropriately labeled dataset. However, the reliance on labeled data is their weakness, and they will not perform well if attack types are unknown to them[2].

Unsupervised Learning Paradigms

To overcome the limitations of supervised methods, unsupervised learning models like Isolation Forest have also been introduced. Isolation Forest excels at anomaly detection by identifying traffic patterns that deviate significantly from the norm, making it effective for detecting zero-day attacks [3].Unlike supervised models, it is not dependent on labeled datasets and thus can be more adaptive to dynamic network environments. [4] emphasize the scalability of Isolation Forest in discovering new threat vectors without a high overhead of computation.

Hybrid Models

Hybrid approaches combining supervised and unsupervised methods have been highly effective. For example, Random Forest combined with Isolation Forest utilizes the unique strengths of each model-accurate classification of known threats, coupled with adaptive detection of novel anomaly patterns [5][6]. Such hybrid combinations improve detection performances while simultaneously reducing false positives, which have long plagued IDSs[7].

Key Challenges in IDS

False Positives: IDS encounters high false-positive rates, which tend to cause "alert fatigue" among network administrators. To overcome this challenge, anomaly detection thresholds need to be fine-tuned and contextual analysis models must be integrated to reduce unnecessary alarms[4].

Scalability: Real time IDS needs to handle high-volume network traffic efficiently. Ensemble methods like Random Forest with light-weight isolation forests as anomaly detectors ensure scalability without sacrificing precision[8],

Dynamic Threats: With time, the nature of cyberattacks changes. Hence, the IDS needs to upgrade their patterns. Hybrid models and continuous learning frameworks ensure that systems function effectively in a dynamic environment[3]

Advancements in Research
The new findings in IDS research emphasize the increase in interpretability of the model by using XAI for improving trust and comprehensiveness of ML models when applied to the critical applications. Additionally, the comparison of supervised and unsupervised methods shows that the right choice of model is strictly bound by scenarios such as high-speed enterprise networks or the resource-limited IoT system [5][7].

Machine learning significantly enhances IDS capabilities, addressing the limitations of traditional systems while offering robust detection mechanisms for both known and unknown threats. Combining the precision of Random Forest with the adaptability of Isolation Forest represents a promising direction for future IDS implementations. Further research into hybrid models, real-time scalability, and explainable frameworks will ensure the continued evolution of IDS in the face of ever-changing cyber threats.

III. METHODOLOGY

Our IDS design was accomplished using our application of trained machine learning models on the NSL-KDD dataset. This dataset contains 41 features in it, including connection duration, protocol type, and any service accessed, in addition to attack class labels, such as 'normal', 'DoS', 'Probe', etc.

3.1 Data Preprocessing

Before training the machine learning models, we first preprocessed the dataset for applicability to analysis:

One-Hot Encoding: All the categorical variables, such as `protocol_type`, `service,` and `flag,` are encoded using one-hot encoding.

Normalization had required normalizing numerical features to guarantee all values were on comparable scales, thus improving the capability of a model.

Dimensionality reduction with the application of the PCA technique would reduce the computation time as follows.

3.2 Model Selection

There were selected two machine learning models for this study:

Isolation Forest is an unsupervised learning algorithm designed to detect anomalies in network traffic by isolating outliers a supervised learning algorithm called Random Forest is used to classify network traffic as normal or attack. Both models were trained using the preprocessed NSL-KDD dataset, with a train-test split ratio of 80:20, as for validation.

3.3 Training and assessment All the models are trained by cross-validation application for robustness and performance appraisals based on accuracy, precision, recall, and F1-score

IV. IMPLEMENTATION

This implementation includes an Intrusion Detection System (IDS) developed with a combination of machine learning along with signature-based detection methods for malicious activities in the network. The system is developed using Python and has a graphical user interface (GUI) devised with the Tkinter library, thus making it accessible and user-friendly for real-time network monitoring and security analysis. The system can detect established attack patterns by predefined signatures, as well as unidentified threats, by identifying anomalies in the network traffic through machine learning models.

The architecture of the system comprises several components, such as data collection, feature extraction, signature-based detection, and anomaly detection based on machine learning. Network traffic is captured through the utilization of the Scapy library, which facilitates real-time packet sniffing. As packets are captured, the Intrusion Detection System copies specific features like size in packet, src bytes size, duration, and other TCP/IP related features for further detailed analysis. It uses signature-based detection to identify the pre-known malicious patterns of established packet payloads; if some kind of match, for example, a particular string string, like "malicious_string_1" or "malicious_string_2", occurs, the IDS alerts the user right away. It has, therefore, implemented a machine learning algorithm called Isolation Forest for anomaly detection in network behavior indicative of the onset of a zero-day attack.

Such a system is designed to support the loading and preprocessing of multiple CSV datasets encompassing network traffic data. Merging of multiple files into a unified dataset is left to the choice of the users and then normalized through Min-Max scaling, which maintains consistency in data. The trained model is isolated with regards to this preprocessed data. The contamination factor has been set at 10% that would enable the outlying cases. The trained model and the scaler are saved using the joblib library to be used in later detection sessions so the model does not need to be trained afresh for every session of detection.

Real-time packet capture and analysis form the core of IDS functionality.

The system uses the sniffing capability of Scapy to capture live network traffic over a specified time range. Each packet then gets analyzed with dual analysis: signature-based detection of known threats and machine learning-based anomaly detection for previously unknown patterns. Features extracted from each packet get processed, and then with the Isolation Forest model, they are classified as either normal or anomalous.

These packets are detected as anomalous and analyzed further; users receive alerts through pop-ups.

The graphical user interface greatly increases the interaction with the Intrusion Detection System as it uses clickable buttons to get access to so many functionalities. Such functionalities include options ranging from starting packet sniffing, find anomalies, load data sets, train a model, clear captured traffic logs, and exit the application. A Text widget can be used to display information relating to captured traffic, such as source and destination IP addresses, packet lengths, and timestamps. Alerts are displayed as message boxes, which allow users to have direct feedback about potential threats or the system itself. The system makes use of multithreading to ensure that a readable graphical user interface is available for packet sniffing and anomaly detection. This makes the operations possible in the background even when using the application with no glitches and freeze. It is because this methodology makes the Intrusion Detection System (IDS) easy to use with fine performances and hence applicable to real-time network security monitoring. Conclusion: The Intrusion Detection System integrates both signature-based detection of known threats and machine learning-based anomaly detection to identify unknown attacks, providing a robust solution to detect network intrusion.Also, the system brought in real-time monitoring and an accessible interface but with the capability of automated machine learning into complete security enhancement tools for networks. Some of the future improvements include further feature engineering, experimentation with alternative models of machine learning, databases for the storage of traffic logs, and further advanced alerting mechanisms such as email or SMS notifications.

V. RESULT

The Integrated Intrusion Detection System combined signature-based and machine learning techniques for detecting harmful activities in the network. It captures live network traffic with Scapy and extracts source IP, destination IP, and packet size as features. Included is a GUI with Tkinter for users to start sniffing packets, train models, and detect anomalies.

Focal Outputs:

Signature Detection: Signature-based IDS was efficient in knowing the threat by looking up packet payloads upon its predefined malicious signatures. On detection of signatures such as malicious_string_1, alerts popped up immediately.

Anomaly Detection: Using an Isolation Forest method, model training for detecting anomalous packets was conducted on normalized network data. This approach runs outlier detection with success, shining light on unknown-types of attacks which lack predefined signatures.

Model Training and Performance: The IDS provides loading of multiple datasets and pre-processing them for model training. For improving detection accuracy, the system applies min-max scaling. Trained models were saved for better reusability.

Responsive GUI: The use of read-write threading guaranteed that the GUI would remain responsive while packets were being captured and investigated in real-time, significantly improving the user's interaction without performance lags.

Overall, the IDS combines signature detection for known threats and machine learning detection of zero-day attacks. Thus, such a system would offer a promising solution for real-time security in networks. Sender may propose to integrate better feature extraction methods, other machine-learning models, and emergency alert notification systems.

V. CONCLUSION

The Intrusion Detection System (IDS) developed in this project successfully combines signature-based detection with machine learning-based anomaly detection to enhance network security. By using real-time packet sniffing and analysis, the system can effectively identify both known threats and zero-day attacks. The use of an Isolation Forest model allows for dynamic detection of anomalies, while predefined signatures ensure immediate alerts for recognized malicious patterns. The integration of a user-friendly graphical interface enhances the system's usability, making it accessible for real-time monitoring and security management. Additionally, the multithreaded architecture ensures that resource-intensive tasks do not compromise the responsiveness of the user interface. Overall, this IDS offers a comprehensive and robust solution for network intrusion detection. Future improvements could focus on expanding feature sets, exploring additional machine learning algorithms, and implementing advanced alerting mechanisms to further strengthen its capabilities. The system demonstrates a promising approach for securing network environments against evolving cyber threats.

VI. REFERENCE

[1] Wang et al., "Machine Learning-Based Intrusion Detection System," IEEE, 2020.

[2] Sharma et al., "A Comparative Study of Machine Learning Methods for Intrusion Detection," IEEE, 2023.

[3] Zhang et al., "An Explainable Machine Learning Framework for Intrusion Detection Systems," IEEE, 2020.

[4] Patel et al., "Addressing Zero-Day Attacks Using Isolation Forest in IDS," IEEE, 2022.

[5] Bala et al., "Comparative Analysis of Intrusion Detection Systems and Machine Learning-Based Model Analysis Through Decision Tree," IEEE, 2024.

[6] Chen et al., "Scalable ML Approaches for IDS in Large Networks," IEEE, 2023.

[7] Gupta et al., "Reducing False Positives in IDS with Machine Learning Models," IEEE, 2024.

[8] IEEE Xplore, "Intrusion Detection Using Machine Learning Techniques: An Experimental Comparison," 2024.