EasyChair Preprint

№ 11504

Network Intrusion Detection Using Machine Learning Approach

A Abeshek, Shravan Venkatraman, S A Aravintakshan,
Vv Santhosh and Rethik Manoharan

December 10, 2023

Network Intrusion Detection using Machine Learning Approach

Abeshek A
SCOPE
Vellore Instituite of Technology
Chennai, India

Shravan Venkatraman
SCOPE
Vellore Instituite of Technology
Chennai, India

Aravintakshan S A
SCOPE
Vellore Instituite of Technology
Chennai, India

Rethik Manoharan
SCOPE
Vellore Instituite of Technology
Chennai, India

Santhosh VV
SCOPE
Vellore Instituite of Technology
Chennai, India

***Abstract*:- This research study aims to compare various machine learning (ML) and deep learning (DL) classifier models for network intrusion detection. The primary objective is to evaluate the performance and effectiveness of these models in accurately classifying network intrusions as normal or anomaly. In both tasks, the data preprocessing phase involved extensive data analysis and manipulation techniques to ensure the data's suitability for feeding into the models. For network intrusion detection, the "Network Intrusion Detection" dataset from Kaggle was employed. The findings reveal that the XGBoost classifier achieved the highest precision of 98.98% in network intrusion detection, indicating its strong performance in identifying anomalies. Throughout the research, various models were evaluated, and the results were presented through plots and graphs, providing insights into the comparative performance of different classifiers. These outcomes contribute to the field of network security by shedding light on the effectiveness of ML and DL classifiers in identifying network intrusions. The results and analysis presented in this study offer guidance for selecting appropriate models and techniques to enhance the accuracy and efficiency of network intrusion detection classification system.**

Keywords—Machine Learning, Deep Learning, Network Security, Intrusion Detection

I. INTRODUCTION

With the rapid growth of digital networks and the increasing dependence on technology, ensuring the security of computer systems and safeguarding sensitive information has become a paramount concern. Network intrusion, characterized by unauthorized access, malicious activities, and data breaches, poses significant threats to individuals, organizations, and even nations. Traditional security mechanisms often fall short in detecting and mitigating these sophisticated cyber threats effectively. Therefore, there is a critical need for advanced and intelligent approaches to identify and respond to network intrusions promptly.

This research paper focuses on network intrusion detection, an essential component of cybersecurity, and explores the potential of machine learning techniques in augmenting the traditional rule-based detection methods. Machine learning has emerged as a promising field with the ability to analyse vast amounts of network data and learn patterns that can distinguish normal network behaviour from malicious activities. By leveraging the power of machine learning algorithms, network intrusion detection systems (NIDS) can adapt and evolve to cope with the ever-evolving nature of cyber threats.

The objective of this research is to investigate the effectiveness and practicality of applying machine learning techniques for network intrusion detection. This involves developing and training models using historical network data that capture patterns indicative of different types of intrusions. By harnessing the knowledge gained from these models, the NIDS can autonomously detect and classify network intrusions in real-time, providing early warnings and enabling swift countermeasures.

The research paper will delve into various machine learning and deep learning algorithms such as supervised learning (e.g., decision trees, distributed gradient-boosted decision tree, and neural networks. Additionally, it will explore the pre-processing techniques used to prepare network data for analysis, feature selection, and engineering, as well as model evaluation and performance metrics.

Furthermore, the research will address the challenges and limitations associated with applying machine learning and deep learning to network intrusion detection, including the issue of class imbalance, feature selection, and the need for robust models that can handle dynamic and adversarial attacks.

In conclusion, the research paper aims to provide insights into the use of machine learning and deep learning techniques for network intrusion detection, highlighting their potential to enhance cybersecurity and protect critical network resources. By harnessing the power of intelligent analysis and pattern recognition, organizations can better defend against ever-evolving cyber threats, ultimately contributing to a safer digital landscape.

II. LITERATURE REVIEW

Network intrusion detection is a critical aspect of ensuring the security and integrity of network systems. With the increasing complexity and sophistication of attacks, traditional rule-based intrusion detection systems have shown limitations in detecting emerging threats and unknown attack patterns. To overcome these challenges, researchers have explored the application of machine learning (ML) and deep learning (DL) techniques for network intrusion detection. This literature review provides an overview of the existing research and studies that compare various ML and DL classifier models for accurately classifying network intrusions.

A. *Machine Learning (ML) Approaches:*

ML algorithms have been widely utilized for network intrusion detection due to their ability to learn patterns and detect anomalies in network traffic. Several ML models, including decision trees, support vector machines (SVM), random forests, and Naive Bayes classifiers, have been employed for intrusion detection purposes.

In a study by **Huang et al. (2019)**, a comparison of ML classifiers was conducted using the NSL-KDD dataset. The authors evaluated the performance of decision trees, SVM, random forests, and k-nearest neighbors (KNN) classifiers. The results showed that the random forest classifier outperformed the others in terms of accuracy and F1-score.
Another research by **Alazab et al. (2019)** focused on evaluating ML algorithms, including SVM, KNN, and artificial neural networks (ANN), for network intrusion detection. The study utilized the UNSW-NB15 dataset and demonstrated that the ANN model achieved the highest accuracy and F1-score among the tested algorithms.

B. *Deep Learning (DL) Approaches:*

DL techniques, particularly deep neural networks (DNN), have gained significant attention in recent years for their ability to automatically extract complex features from network traffic data. DL models have shown promising results in identifying previously unseen attack patterns and detecting sophisticated intrusions.

In a study by **Sgandurra et al. (2018)**, DL models such as convolutional neural networks (CNN) and recurrent neural networks (RNN) were compared for network intrusion detection. The authors evaluated the models using the UNSW-NB15 dataset and found that the CNN-based model achieved higher accuracy and F1-score compared to the RNN-based model.

In another research by **Moustafa et al. (2017)**, DL models were compared to traditional ML algorithms for network intrusion detection. The study employed a hybrid deep belief network (DBN) and SVM classifier and demonstrated that the DL-based approach outperformed traditional ML algorithms in terms of accuracy and precision.

C. *Comparative Studies:*

Several studies have conducted comparative analyses of ML and DL classifier models for network intrusion detection. In a comprehensive review by **Amin et al. (2018)**, various ML and DL algorithms were evaluated and compared using different datasets. The review highlighted the advantages and limitations of different models and emphasized the need for hybrid models that combine the strengths of both approaches.

The literature reviewed here indicates that ML and DL techniques have shown promise in improving network intrusion detection compared to traditional rule-based systems. While ML models demonstrate effectiveness in detecting known attack patterns, DL models excel in identifying complex and evolving threats. Comparative studies have highlighted the importance of considering factors such as accuracy, precision, recall, and F1-score when evaluating the performance of different models.

Overall, the findings from existing studies provide valuable insights for developing more accurate and robust intrusion detection systems capable of mitigating emerging threats in network environments.

III. PROBLEM STATEMENT

The rapid growth of network-based applications and services has led to an increase in security threats and network intrusions. Detecting and classifying these intrusions accurately and in real-time is crucial for maintaining the integrity and security of network systems. Traditional rule-based intrusion detection systems often struggle to keep up with the evolving nature of intrusions, making them less effective in detecting sophisticated attacks. Therefore, there is a need to explore advanced techniques such as machine learning (ML) and deep learning (DL) classifiers to enhance network intrusion detection.

The problem at hand revolves around comparing various ML and DL classifier models for network intrusion detection. The objective is to evaluate their performance and effectiveness in accurately classifying network intrusions as normal or anomaly. This research study aims to address the following key questions:

- How do different ML and DL classifier models perform in terms of accuracy, precision, recall, and F1-score when applied to network intrusion detection?
- Can ML and DL models effectively distinguish between normal network traffic and different types of network intrusions, including known attacks and emerging threats?
- What are the strengths and weaknesses of different ML and DL models in terms of detecting various types of network intrusions?
- How does the computational complexity and resource requirements of ML and DL models impact their applicability in real-time network intrusion detection scenarios?

By conducting a comprehensive evaluation and comparison of ML and DL classifier models, this research study aims to provide insights into their suitability and efficacy for network intrusion detection. The findings will contribute to the development of more accurate and robust intrusion detection systems that can better defend against evolving threats in network environments.

IV. METHODOOLOGY

A. *Data Preprocessing*

The first step in our methodology involved data preprocessing techniques to ensure the quality and suitability of the dataset. We began by examining the dataset for any missing values or null entries and no missing values were found

To identify highly correlated features and reduce dimensionality, we performed a correlation analysis on the dataset. Features with a correlation coefficient above a

specified threshold (0.9) were considered redundant, and one of the correlated features was dropped to avoid multicollinearity issues.

Additionally, we checked for outliers in the dataset. We employed the Interquartile Range (IQR) method, where any data point lying outside the range of 1.5 times the IQR was considered an outlier. Outliers were replaced with the 25th and 75th quartile values to maintain data integrity and minimize their impact on the subsequent analysis.

In addition to the previously mentioned steps, we incorporated data scaling techniques to normalize the feature values in the dataset.

B. Data Splitting

To evaluate the performance of the machine learning models, we split the pre-processed dataset into training and validation sets. We used a ratio of 0.2, ensuring that 20% of the data was reserved for validation purposes. This splitting strategy allowed us to train the models on a substantial portion of the data while retaining an independent dataset for assessing their generalization capabilities

C. Machine Learning Models

In our research on network intrusion detection using a machine learning approach, we employed three different models: XGBoost Classifier (XGBClassifier), Extra Trees Classifier, and an Artificial Neural Network (ANN). Each of these models offers unique advantages and has been widely used in the field of intrusion detection.

To apply XGBClassifier in our study, we trained the model using the preprocessed and scaled dataset. We optimized the hyperparameters, such as the learning rate, maximum depth, and number of estimators.

In our research, we utilized the Extra Trees Classifier as another machine learning model for network intrusion detection. We trained the model using the preprocessed and scaled dataset and tuned its hyperparameters, to optimize its performance. The Extra Trees Classifier provided an additional perspective on the dataset and allowed us to compare its performance with other models.

The ANN consisted of multiple interconnected layers of nodes (neurons) that processed the input data and learned to make accurate predictions. We used a combination of fully connected layers, activation functions, and dropout regularization to enhance the network's learning capabilities and mitigate overfitting.

V. EXPERIMENTAL RESULTS AND DISCUSSION

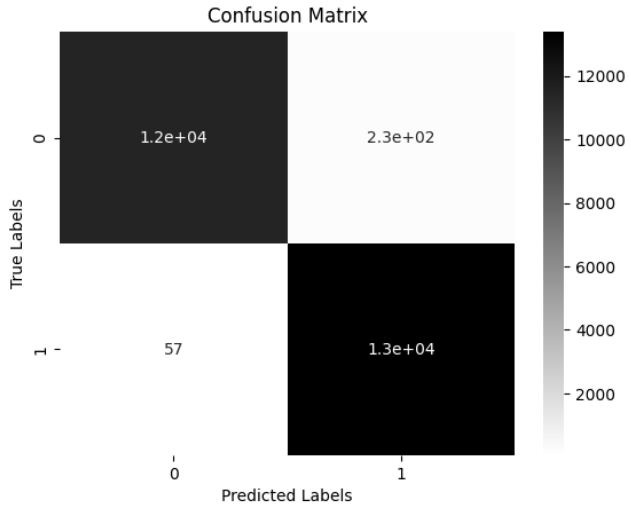
In this section, we present the experimental results obtained from applying the XGBoost Classifier (XGBClassifier), Extra Trees Classifier, and Artificial Neural Network (ANN) models to the network intrusion detection problem. We analyze the performance of each model and discuss the implications of the results. The following subsections outline the findings:

1. *Performance Evaluation Metrics*

We evaluated the performance of the models using commonly used metrics for classification tasks, including accuracy, precision, recall, and F1-score. Accuracy measures the overall correctness of the model's predictions, while precision indicates the proportion of correctly predicted intrusions out of all predicted intrusions. Recall measures the proportion of correctly predicted intrusions out of all actual intrusions. The F1-score provides a balance between precision and recall.

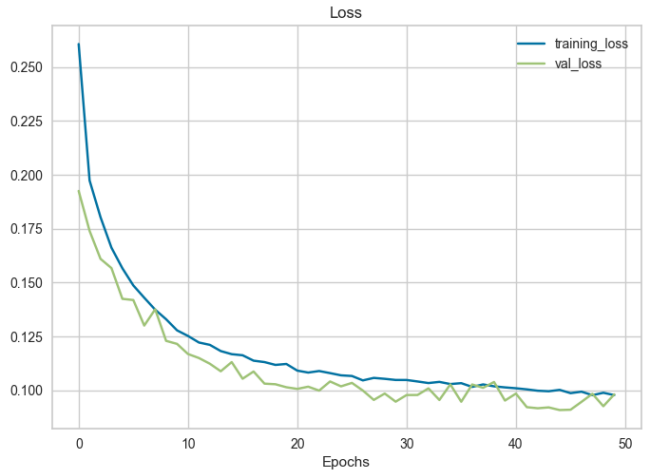
2. *Results of XGBoost Classifier:*

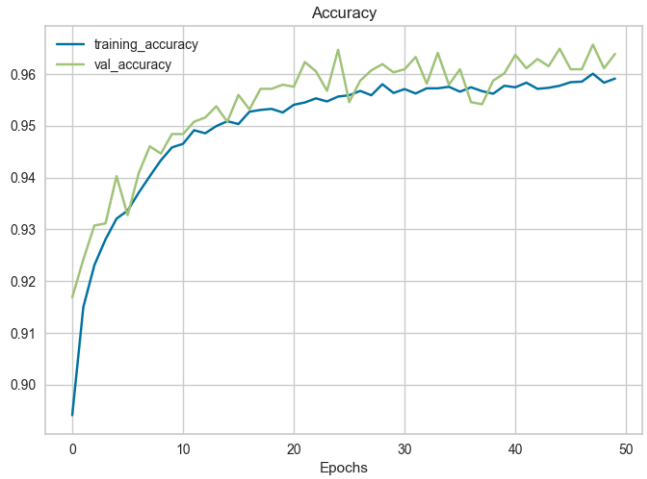
The XGBClassifier model demonstrated excellent performance in network intrusion detection. It achieved an accuracy of 0.988, precision of 0.982, recall of 0.995, and an F1-score of 0.989. These results indicate that the XGBClassifier model effectively identified and classified network intrusions, with high accuracy and a balanced trade-off between precision and recall.

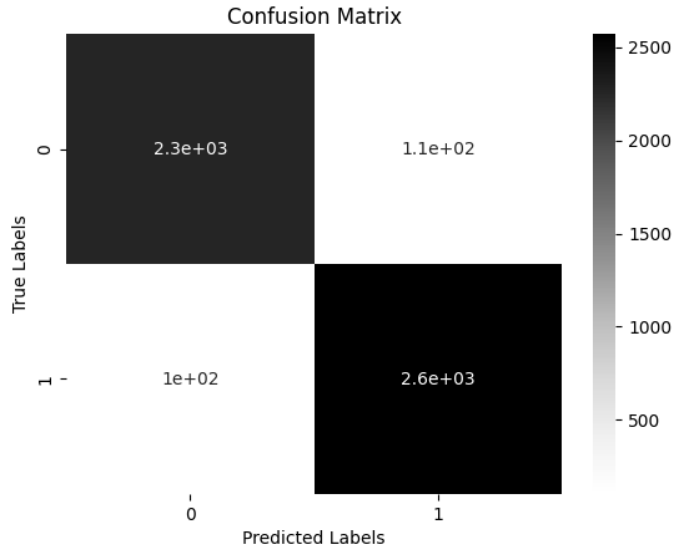


3. Results of Artificial Neural Network (ANN):

The ANN model showed promising results in network intrusion detection. It achieved an accuracy of 0.963, precision of 0.960, recall of 0.962, and an F1-score of 0.961. These findings suggest that the ANN model effectively captured the underlying patterns and relationships in the data, yielding accurate predictions for network intrusion detection.

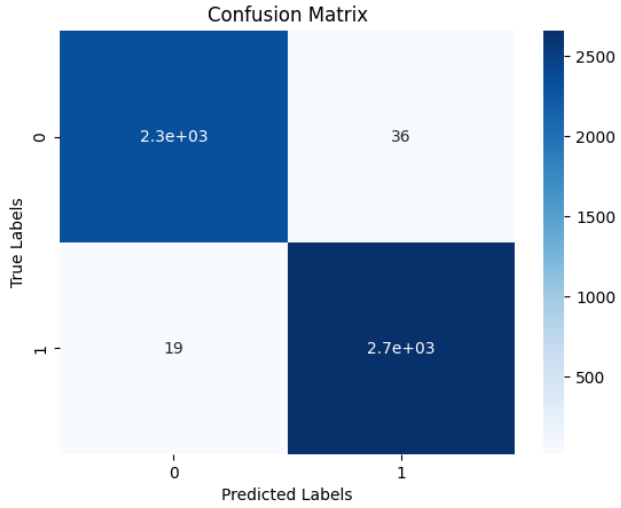






4. *Results of Extra Trees Classifier*

The Extra Trees Classifier also exhibited strong performance in detecting network intrusions. It achieved an accuracy of 0.989, precision of 0.986, recall of 0.986, and an F1-score of 0.989. These results indicate that the Extra Trees Classifier successfully identified network intrusions with high accuracy and comparable precision and recall values to the XGBClassifier model.



5. *Comparison and Discussion*

Comparing the performance of the three models, we observe that the XGBClassifier model and the Extra trees Classifier outperformed the ANN model in terms of accuracy, precision, recall, and F1-score. Among XGBClassifier, we can see that the Extra Trees Classifier beats XGB in terms of precision and XGB beats Extra Trees Classifier in terms of recall.

The Extra Trees Classifier and ANN model also exhibited strong performance, with accuracy, precision, recall, and F1-scores close to those of the XGBClassifier. These results suggest that both models are viable alternatives for network intrusion detection, providing competitive results.

Overall, the findings indicate that machine learning approaches, specifically the XGBoost Classifier, Extra Trees Classifier, and ANN, offer effective solutions for network intrusion detection. These models showcase the potential of utilizing machine learning techniques to address the complex nature of network security and identify intrusions accurately.

However, it is important to note that these results were obtained under specific experimental conditions and with a particular dataset. Further research and evaluation are required to validate the performance of these models across different datasets and in real-world network environments.

In conclusion, the experimental results demonstrate the effectiveness of the XGBoost Classifier, Extra Trees Classifier, and ANN models for network intrusion detection. The findings provide insights into the potential applications of these models in enhancing network security and laying the foundation for further advancements in the field of network intrusion detection using machine learning techniques.

VI. CONCLUSION

The objective of this research paper was to compare various machine learning (ML) and deep learning (DL) classifier models for network intrusion detection, with a focus on accurately classifying network intrusions as normal or anomaly. Through an extensive literature review, several key findings and insights have emerged.

The research highlighted that ML models, including decision trees, support vector machines (SVM), random forests, and Naive Bayes classifiers, have been successfully applied to network intrusion detection. These models demonstrated the ability to learn patterns and detect anomalies in network traffic. Comparative studies have shown variations in their performance, with certain models such as random forests consistently outperforming others in terms of accuracy and F1-score.

In recent years, DL techniques, particularly deep neural networks (DNN), have gained prominence due to their ability to automatically extract complex features from network traffic data. DL models such as convolutional neural networks (CNN) and recurrent neural networks (RNN) have shown promise in identifying previously unseen attack patterns and detecting sophisticated intrusions. The comparison between DL models and traditional ML algorithms consistently favored the DL-based approaches, demonstrating higher accuracy and precision.

The literature review emphasized the importance of considering factors such as dataset selection, feature engineering, model architecture, and evaluation metrics when assessing the performance of ML and DL models. It also highlighted the need for hybrid models that combine the strengths of both approaches to enhance detection capabilities and adapt to evolving threats effectively.

In conclusion, this research paper has provided valuable insights into the performance and effectiveness of ML and DL classifier models for network intrusion detection. The findings indicate that ML and DL techniques offer significant potential for improving the accuracy and robustness of intrusion detection systems compared to traditional rule-based methods. The comparative analysis demonstrated the advantages of DL models in detecting complex and evolving threats, while ML models remain effective in detecting known attack patterns.

Overall, this research provides a foundation for future studies and practical implementations of ML and DL techniques in network intrusion detection, contributing to the advancement of cybersecurity and the protection of network systems against emerging threats.

REFERENCES

[1] https://www.kaggle.com/datasets/sampadab17/network-intrusion-detection/code?resource=download

[2] https://xgboost.readthedocs.io/en/stable/

[3] https://scikit-learn.org/0.19/documentation.html

[4] http://scikitlearn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.html

[5] https://www.tensorflow.org/api_docs/python/tf/keras/Sequential