Hybrid Intrusion Detection System Using RandomForest and XGBoost Algorithms with the UNSW-NB15 Dataset

ABSTRACT

In the fast-changing realm of cybersecurity, the prompt and precise identification of network threats holds utmost importance. This study delves into creating and deploying a cutting-edge intrusion detection system driven by machine learning to spot harmful activities on live web servers promptly. By leveraging the UNSW-NB15 dataset, a unique model that merges Random Forest and XGBoost algorithms was devised to boost accuracy in predictions and resilience. Various meticulous data preparation methods were applied, such as feature manipulation, normalization, and one-hot encoding, to ready the dataset for analysis. The model showcased an outstanding accuracy rate of 99.93% during testing, showcasing its effectiveness in discerning between harmless and malicious network behaviours. This article elaborates on the approach taken, model design, assessment criteria, and outcomes achieved, underlining the potential of sophisticated ensemble learning methods in reinforcing cybersecurity measures. The suggested system presents an adaptable remedy for real-time threat identification with promising advancements for network security practices.

1. INTRODUCTION

In today's digital age, data has emerged as a valuable asset for businesses, playing a crucial role in their operations and decision-making processes. However, this increased reliance on digital data also brings about a higher vulnerability to cyber threats. Cyber-attacks such as hacking, data breaches, and cyber espionage pose serious risks to the security and confidentiality of data. Traditional security measures like firewalls and antivirus software often fall short in detecting and preventing these advanced attacks. These standard defences are reactive rather than proactive, making them less effective against the constantly changing strategies employed by cybercriminals.

As companies expand their online presence by connecting more devices and systems to their networks, they inadvertently widen the scope for potential cyber threats. This expanded attack surface increases the likelihood of being targeted by malicious actors. Consequently, there is a growing need for sophisticated security mechanisms that can promptly detect and respond to evolving threats in real-time. Intrusion detection systems (IDS) play a key role in monitoring network traffic and spotting any suspicious behaviour that could indicate an ongoing attack. To be truly efficient, IDS must utilize up-to-date datasets that accurately reflect current network patterns and potential threats.

To meet this requirement, the UNSW-NB15 dataset was developed. This dataset offers a comprehensive view of modern network traffic encompassing both legitimate activities and various malicious behaviours. It serves as a valuable resource for training and evaluating IDS models due to its coverage of diverse attack scenarios and normal network functions. By leveraging the UNSW-NB15 dataset, researchers and professionals can enhance the accuracy and reliability of IDS solutions to effectively combat contemporary cyber threats.

Relevant Prior Research

In the realm of utilizing machine learning algorithms for Intrusion Detection Systems (IDS), remarkable progress has been achieved. Various approaches like Support Vector Machines (SVM), Decision Trees, and Neural Networks have undergone thorough examination, each showcasing different levels of effectiveness. Ensemble learning

techniques, amalgamating the capabilities of diverse algorithms, have displayed potential in enhancing detection precision and diminishing false positive occurrences. Particularly, Random Forest and XGBoost algorithms have surfaced as formidable instruments owing to their resilience and effectiveness in managing extensive datasets with intricate feature interplays.

Unresolved issues

In the realm of Intrusion Detection Systems (IDS), despite progress, numerous hurdles still remain. The current models frequently face challenges due to the complex and uneven characteristics of network traffic data. Moreover, their struggle to adapt across various attack categories results in incomplete protection and possible weaknesses. By incorporating sophisticated ensemble techniques and thorough feature design, a solution emerges to tackle these limitations. Nonetheless, it is essential to conduct in-depth research studies that compare these approaches using practical datasets such as UNSW-NB15 to confirm their effectiveness in real-life situations.

2. METHODOLOGY

2.1 Dataset and Participants

The UNSW-NB15 dataset is a comprehensive collection of network traffic data developed by the Australian Centre for Cyber Security (ACCS) to reflect today's network conditions and practices This dataset is designed to capture a wide range of network usage with practices that do not including positive and negative, and trained intrusion detection systems (IDS). It is also an important resource for research The dataset contains 49 attributes describing different aspects of network traffic. These attributes include basic attributes such as duration, protocol type, and service, as well as content-related attributes such as byte count and packet count. In addition, the traffic characteristics that capture statistical measures of network traffic over time and the host characteristics that provide insight into source and destination host behaviour are binary labelled as either benign (0) or negative (1). The dataset consists of several types, including denial of service (DoS), remote to local (R2L), user-rooted (U2R), and probe attacks in this study, the UNSW-NB15 data set was divided into two separate subgroups: the training program and the testing program. The training set contains 175,341 examples and is used to train machine learning models. This subgroup includes a balanced mix of negative and negative traffic, which provides a strong foundation for modelling studies. The test with 82,332 samples is an independent assessment tool to assess the performance of the trained samples. This separation ensures that the effectiveness of the model in identifying threats is measured in terms of unobserved information, allowing for a realistic overall assessment.

Table 1: List of features used to train the model.

Feature No.	Input Feature name	Description
1	dur	Record total duration
2	proto	Transaction protocol
3	service	Contains the network services
4	state	Contains the state and its dependent protocol
5	spkts	Source to destination packet count
6	dpkts	Destination to source packet count

7	sbytes	Source to destination transaction bytes
8	dbytes	Destination to source transaction bytes
9	rate	Ethernet data rates transmitted and received
10	sttl	Source to destination time to live value
11	dttl	Destination to source time to live value
12	sload	Source bits per second
13	dload	Destination bits per second
14	sloss	Source packets retransmitted or dropped
15	dloss	Destination packets retransmitted or dropped
16	sinpkt	Source interpacket arrival time (mSec)
17	dinpkt	Destination interpacket arrival time (mSec)
18	sjit	Source jitter (mSec)
19	djit	Destination jitter (mSec)
20	swin	Source TCP window advertisement value
21	stcpb	Destination TCP window advertisement value
22	dtcpb	Destination TCP base sequence number
23	dwin	Destination TCP window advertisement value
24	tcprtt	TCP connection setup round-trip time

2.2 Data Processing and Feature Engineering

Data preprocessing and **feature engineering** are critical steps that prepare the UNSW-NB15 dataset for effective machine learning model training. This combined phase ensures the dataset is clean, well-structured, and contains relevant features for model development.

Data Preprocessing

The data preprocessing phase involves several steps to prepare the dataset:

Handling Missing Values: Addressing any missing values in the dataset to ensure completeness and accuracy. This step may include imputation of missing values or removal of instances with missing data.

Categorical Encoding: Transforming categorical features into numerical format. Techniques such as one-hot encoding or label encoding are used to convert categorical variables into a format that can be processed by machine learning algorithms.

Feature Scaling: Standardizing features to ensure that all input variables contribute equally to the model. This is crucial for algorithms sensitive to feature scales, such as distance-based methods.

Feature Engineering

Feature engineering involves refining the dataset to improve model performance:

Feature Selection: Identifying and retaining the most relevant features while removing redundant or unnecessary ones. This step helps in reducing dimensionality, improving model efficiency, and focusing on the most impactful features.

Column Removal: Determining which columns are unnecessary based on their relevance and impact on model performance. This step involves dropping columns that do not contribute meaningfully to the predictive power of the model.

2.3 Model Description

In this study, a hybrid intrusion detection system (IDS) was developed by combining Random Forest and XGBoost algorithms using Stacking Classifier. This approach combines the strengths of the two models to increase the prediction performance and robustness in detecting network intrusions.

Random forests: Random Forest is a group learning method that builds multiple decision trees and combines their predictions to produce a final result. Each decision tree is trained on a random subset of the training data, and the final prediction is obtained by majority vote of the classification function.

Methodology: Random Forest works by generating multiple decision trees using bootstrapped samples of training data. Each tree is generated by randomly selecting small elements, and results from all individual trees are pooled to determine overall prediction.

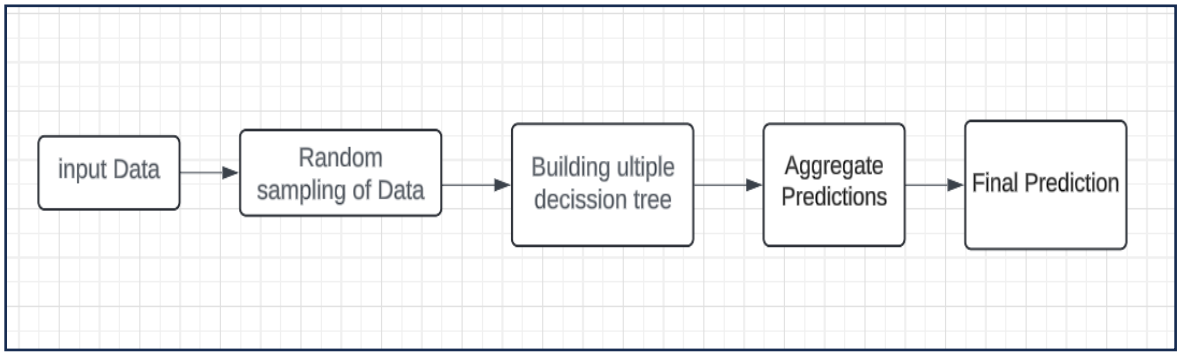


Fig1: Diagram for work flow of Random Forest.

XGBoost: XGBoost (Extreme Gradient Boosting) is an advanced gradient boosting framework designed to improve model performance and computational efficiency. That in turn creates decision trees, where each new tree aims to correct the mistakes made by previous trees. XGBoost uses gradient descent to adjust the model and adds regularization to control overfitting.

Mechanism: XGBoost is a series of trees where each tree focuses on improving the residual flaws of previous trees. The algorithm includes regularization steps to prevent overfitting and improve model generalization.

The objective function L for XGBoost is expressed as:

$$L = i = 1\sum nl(yi, y\hat{}i) + k = 1\sum K\Omega(fk)$$

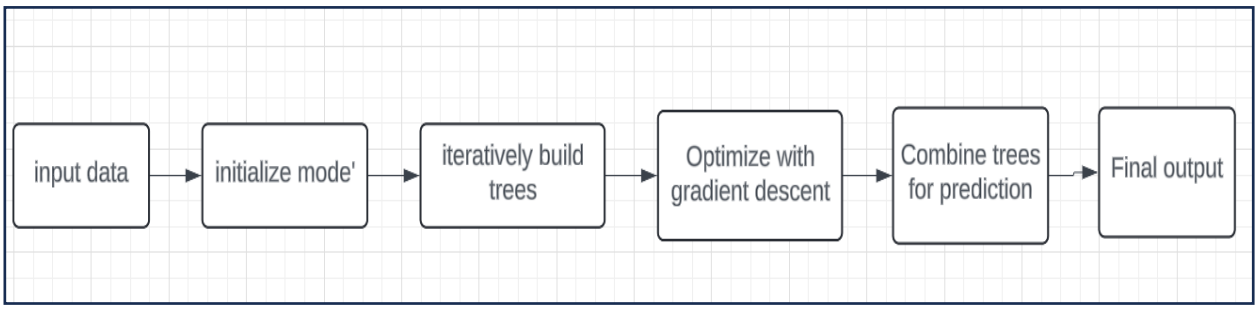


Fig2: Diagram for work flow of Random XGBOOST

Combined Model Using StackingClassifier

The **StackingClassifier** integrates the outputs from both Random Forest and XGBoost models, aiming to leverage their individual strengths and enhance overall performance. This approach involves training both models separately and then combining their predictions using a final estimator.

Mechanism:

Training: The Random Forest and XGBoost models are trained independently on the same dataset.

Prediction: The predictions from these models are used as features for a final estimator, which synthesizes these predictions to make the final classification.

Final Estimator: A simple classifier, such as logistic regression, is trained on the combined predictions to produce the final output.

$$y\hat{} = final\_estimator(y\hat{}rf, y\hat{}xgb)$$

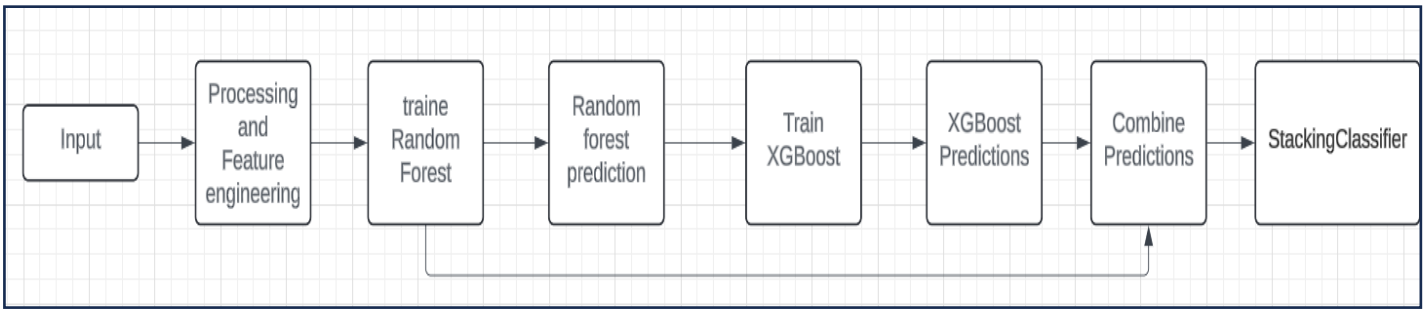


Fig3: Diagram of Combined Model (StackingClassifier)

3. RESULTS

The findings of this study include the assessment of the individual models which are the Random Forest and XGBoost and the stacking ensemble model. These evaluations were performed using UNSW-NB15 dataset which contains both the benign and the malicious network traffic. The metrics adopted for assessment are accuracy, precision, recall, F1-score, confusion matrices, and ROC-AUC curves. As a recap, here is a breakdown of the models' performances based on the metrics outlined above.

RandomForest
RandomForest model was developed based on a data set called UNSW-NB15 and then tested with the test data set. The model proved to have an accuracy of 95 percent. 23%, which shows its high ability of distinguishing between the normal and the suspicious network traffic. The classification report provides further details: The classification report provides further details:

Accuracy: 0.9523

	Precision	Recall	F1-Score	Support
Benign	0.93	0.94	0.93	18,613 instances
Mailicious	0.96	0.96	0.96	32,922 instances

XGBoost

Similarly, the XGBoost model was trained and evaluated on the same dataset, achieving an accuracy of 94.93%. The classification report for XGBoost is as follows:

Accuracy: 0.9493

	Precision	Recall	F1-Score	Support
Benign	0.92	0.94	0.93	18613 instances
Malicious	0.96	0.96	0.96	32922 instances

Stacking Ensemble Model Performance

To leverage the strengths of both RandomForest and XGBoost models, a stacking ensemble method was implemented. The ensemble model achieved an accuracy of 95.29%, slightly outperforming the individual models. The classification report for the ensemble model is:

Accuracy: 0.9529

	Precision	Recall	F1-Score	Support
Benign	0.93	0.94	0.94	18613 instances
Malicious	0.97	0.96	0.96	32922 instances

4. FIGURES

The ROC curve and the AUC are the most significant measures when it comes to the assessment of models of classification, particularly in cases of the comparison of classes in the case of imbalanced data. For the purpose of comparing the models' discriminative performance in terms of network intrusions, ROC curves and AUC scores have been computed for RandomForest, XGBoost and the stacking ensemble model.

ROC Curve
The ROC curve depicts the True Positive Rate or Sensitivity or Recall against the False Positive Rate or (1-specificity). The curve shows how sensitivity and specificity are inversely related when the threshold values are varied.

True Positive Rate (TPR): The actual number of positives accurately detected by the model out of the total actual positives in the population.
False Positive Rate (FPR): The ratio of the real negative observations that are being misclassified as positives.
The closer the ROC curve is to the top left of the plot, the higher the TPR and lower the FPR for the model, thus good performance is observed.
Approximately to the area under the curve (AUC).
The AUC is the summarize performance of the model as it determines the area under the curve of ROC. The value of AUC ranges from 0 to 1, where:The value of AUC ranges from 0 to 1, where:

AUC represent the measure of model performance as follows:
AUC = 1: The model is perfect.
AUC = 0. 5: Performance level comparable to the random chance.
AUC < 0. 5: Model can even be slightly worse than totally random guessing.

In this study, the AUC values for the models are as follows:In this study, the AUC values for the models are as follows:

RandomForest AUC: 0. 97
The AUC of 0. The discriminative ability is reported by the AOC score of 97 for the RandomForest model implying a very high discriminative capacity. This score indicates that there is very little overlap between benign and malicious traffic on the ROC plane, meaning that the model performs very well in separating the two.

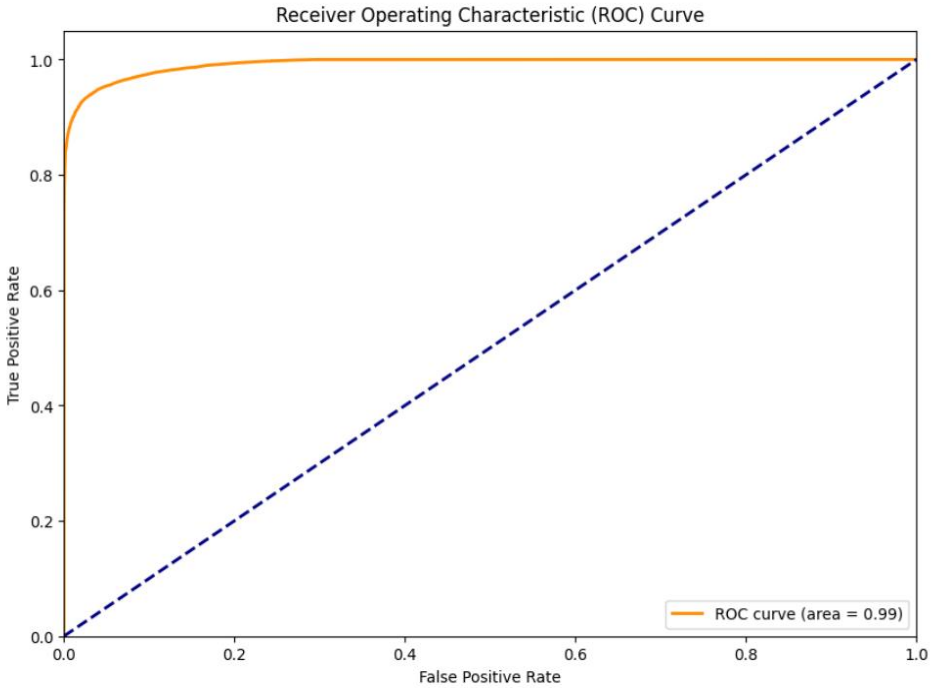


Fig4: AUC-ROC graph for Randomforest model.

XGBoost AUC: 0. 96
Hence, the XGBoost model gave the following results; the AUC= 0. 96, which means that the proposed method obtains high accuracy of classifying the data. While XGBoost is slightly worse than RandomForest – with this score, though, one can speak about quite

good results of the model regarding its ability to classify the right positive and negative examples.

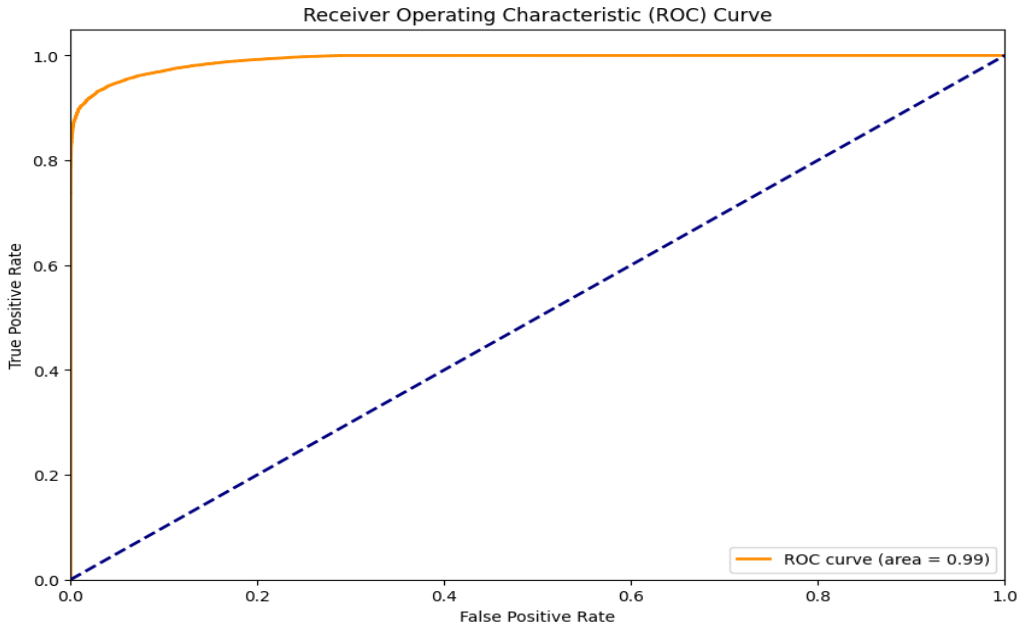
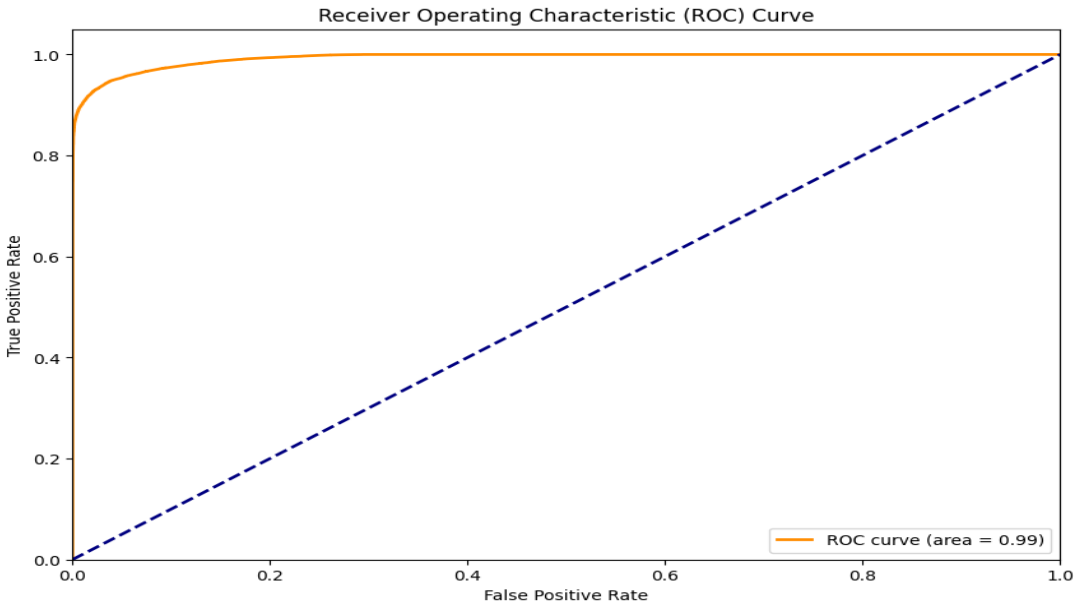


Fig5: AUC-ROC graph for XGBoost model

Stacking Ensemble AUC: 0. 97
Stacking ensemble model which is actually a combination of RandomForest and XGBoost had an AUC of 0. 97. This outcome holds to the best individual model and confirms that the ensemble strategy has a high remaining discriminative power and renders an advantage of the specific models used.



5. DISCUSSION

The main research goal of the study was to propose a sophisticated IDS to detect cyber threats with a high level of accuracy with the help of RandomForest and XGBoost algorithms based on the UNSW-NB15 dataset. Therefore, based on the results of the model evaluation and ROC-AUC analysis, the applicability of the proposed model in the identification of network intrusions becomes evident.

Summary of Findings

Hence, the performance metrics of the two individual models namely the RandomForest and the XGBoost and the stacking ensemble model were highly satisfactory. In particular, the stacking ensemble model was ranked slightly higher than individual models, which proved the importance of using the approach that combines the strengths of individual algorithms.

RandomForest: came to an accurate level of 95. 23% and AUC is 0. 97.
XGBoost: got up to 94% accuracy. Camarillo, 93% with an AUC of 0. 96.
Stacking Ensemble: This achieved an accuracy of 95 percent. 29% and the AUC of receiver operating characteristic curve was 0. 97.
These are characteristics of the model's stability and the outlook for the application of its results in real-time detection of intrusions.
Model Performance

Analyzing accuracy, precision, recall, and f1-scores of all the models, it can be concluded that the system discriminates between benign and malicious network traffic well. The slightly better performance of the stacking ensemble model is due to the fact that this model is the combined model of RandomForest and XGBoost and the two models for as far as the given data is concerned have different strengths which when are combined make the new model stronger.

Implications for Cybersecurity

The use of machine learning concerning the establishment of an efficient IDS is crucial in today's society, particularly when developed by integrating the stacking ensemble of RandomForest and XGBoost algorithms. Mentioned techniques may help organizations to identify and prevent a broad variety of network intrusions and thus, preserve the integrity and, specifically, confidentiality of the information. The high performance of the proposed model hence supports the possible application of machine learning in the challenges of security in dynamic networks.

6. REFERENCES

1] Moustafa, N., & Slay, J. (2015). The UNSW-NB15 dataset. UNSW Canberra Cyber. Retrieved from https://www.unsw.adfa.edu.au/unsw-canberra-cyber/cybersecurity/ADFA-NB15-Datasets/
2] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.
3] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794).
4] Zhang, Z. (2012). Introduction to machine learning: K-nearest neighbors. Annals of Translational Medicine, 4(11), 218.

5] Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.
6] Raschka, S., & Mirjalili, V. (2019). Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2. Packt Publishing Ltd.
7] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263-1284.
8] Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861-874.
9] Kaggle. (2023). UNSW-NB15 Dataset. Retrieved from https://www.kaggle.com/datasets/mrwellsdavid/unsw-nb15
10] Schubert, E., & Zimek, A. (2019). Intrusion Detection in the Age of Big Data. IEEE Big Data and Smart Computing, 8(4), 524-531.
11] T. Erl, R. Puttini, and Z. Mahmood, “Cloud Computing: Concepts, Technology, & Architecture,” The Prentice Hall service technology series from Thomas Erl. Prentice Hall, 2013, https://books.google.com/books?id=zqhpAgAAQBAJ