National College of Ireland

Implementing Machine Learning Algorithms To Enhance Intrusion Detection System Across Computerised Networks Towards Pre-empting Cyber Attacks

MSc Research Project
Cybersecurity

Shruti Praveen Garg
Student ID: x23206047

School of Computing
National College of Ireland

Supervisor: Prof. Liam Mccabe

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Shruti Praveen Garg
Student ID:	X23206047
Programme:	MSc. Cybersecurity **Year:** 2024-2025
Module:	Research Thesis
Supervisor:	Prof. Liam Mccabe
Submission Due Date:	12th December 2024
Project Title:	Implementing Machine Learning Algorithms To Enhance Intrusion Detection Systems Across Computerised Networks Towards Pre-empting Cyber Attacks.
Word Count:	6861 **Page Count:** 23

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Shruti Praveen Garg

Date: 12th December 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Table of Content

Implementing Machine Learning Algorithms To Enhance Intrusion Detection Systems Across Computerised Networks Towards Pre-empting Cyber Attacks.

Shruti Praveen Garg
X23206047

Abstract

The development of cyber attacks has significant threat on the security of any networked system. This research highlights the advancement of machine learning to enhance Intrustion Detection System for detection of unforeseen cyber attacks, detection and mitigation of a variety of network intrusions namely, Distributed Denial of Service Attack, Brute Force, SQL Injection, etc. on corporate network. With the use of CIC-IDS 2017 dataset which resembles real-world data, the study compares the execution of multiple models like Logistic Regression, Random Forest, Isolation Forest, and One-Class SVM individually and combines them into a Voting Classifier using soft voting with minimizing false positives and detecting accuracy being the pivotal part of the scale. Following a sequence of preprocessing steps which included feature scaling, label encoding, and missing data imputation, the dataset was assembled for model training. To understand the relevant characteristics for intrusion detection, feature selection techniques were applied, and the models were trained and assessed using classification metrics, precision, recall, F1-score, and ROC-AUC for an extensive assessment of the model performances. Ensemble mechanism- Voting Classifier was implemented in combining the prediction of all the models to magnify the detection accuracy. The results display that a combination of ensemble-based technique with supervised learning and anomaly detection technique gives more advanced performance in detecting network anomalies when compared with individual models. This study highlights the ability of machine learning approaches in enhancing the reliability and effectiveness of intrusion detection system also offers insights to secure corporate network from advanced cyber-attacks.

Keywords- Intrusion detection system, machine learning, data preprocessing, model, training, classification metrices, precision, recall, F1-score, Logistic Regression, Deep Neural Network, Isolation Forest, One-Class SVM, Random Forest, Voting Classifier.

1 Introduction

There has been a recorded spike in cyber-attacks in the second half of 2024 with a 30% hike globally, going up to 1,636 attacks per week per organization when compared to the same period of 2023(gmcdouga, 2024). Since, organizations highly depend on digital frameworks, the requirement for strong Intrusion Detection Systems has become a serious aspect. Traditional intrusion detection system often uses signature based or rule bases detection methods, which are effective in case of known threats but are weak to detect unknown patterns, for instance zero-day attacks or advanced persistent threats. All these constraints have directed

towards the path of exploring artificial intelligence and machine learning approaches to boost the IDS abilities.

1.1 Motivation

Machine Learning provided the capability to analyse large number of network traffic data and recognise irregularities which might indicate malicious activities. Studies done by researchers previously(Disha and Waheed, 2022) (Tait *et al.*, 2021), in which the application of machine learning techniques in IDS to enhance advance threat classification is explored. Selecting the suitable algorithm is a complex task, and the study(Tait *et al.*, 2021) analyses the current condition of intrusion detection approach and discusses it's advantages and disadvantages. When trained on labelled datasets, supervised learning models, Logistic Regressions and Random Forests work extremely well with known attack patterns, similarly unsupervised models, Isolation Forest, One-Class SVM are proficient at detecting anomalies. Combining multiple algorithms show improvement in detection and accuracy, even then, the main struggle to handle large number of datasets, choosing important features and establishing balance between big detection and low false alarm. To direct these problems, this study works on existing research by combining supervised learning models, anomaly detection, and ensemble learning approaches to enhance Intrusion Detection System focusing on computerised networks for organizations.

1.2 Research Question

How can integrating supervised learning and anomaly detection models in ensemble framework enhance detection accuracy and reliability while minimizing false positives and ensuring scalability for network security?

1.3 Research Overview

This research works with CIC-IDS 2017 dataset which has multiple types of attacks and is used as a standard dataset in intrusion detection system research(*IDS 2017 | Datasets | Research | Canadian Institute for Cybersecurity | UNB*, 3rd July, 2017)). The study worked in a structured pattern to address the problems related to network intrusion detection system. First, the dataset was pre-processed where missing values where handled and infinite values were replaced for data integrity. To bring the features to line StandardScaler, and to convert the target variable in numeric form Label Encoder was used, for feature selection SelectKBest was utilized to recognize top 10 features which are most important for intrusion detection. In this step, the computational efficiency of the model was enhanced, and it also optimized the interpretability. For model development, to classify network traffic as Benign or malicious, on the pre-processed dataset supervised learning models, Logistic Regression and Random Forest were trained and for anomaly detection Isolation Forest and One-Class SVM were executed to detect irregularities which are useful to recognize rare or new attack patterns. To integrate the predictions from all the models, and leveraging their individual robustness to enhance the overall detection performance, Voting Classifier an Ensemble method was used. It uses Soft Voting where the predicted probability by each model is averaged to decide the final prediction. As this method utilizes the strengths of the models used. Moreover, it can also handle imbalanced data while also reducing errors of the individual model. Evaluation of the models

were done using precision which is also known as positive predictive value, it measures the amount of true positive and false positives instances belong to a class. Recall is the matrix which says how many true positives and false positives were rightfully classified. F1-score is the mean of precision and recall whose value lies between 0 and 1(Priyanka, 2022). Confusion matrix is considered as a performance measurement for machine learning(Narkhede, 2021). ROC-AUC score shows model performance at different values used in classification(Priyanka, 2022). These metrics offers a vast understanding of the models effectiveness in detecting intrusions. This study shows an alliance between supervised and unsupervised learnings in intrusion detection systems. It also shows that preprocessing and feature selection have a huge impact in the performance of a model. Ensemble learning approach can surpass independently working models as their strengths are combined, and machine learning offers scalable and compliant results in securing networks against mature cyber threats. The importance of combining multiple machine learning techniques in building a strong intrusion detection system for real world implementations is highlighted with the results of this study. This study not just confirms the applicability of already existing studies but also provides a way for further development in this region.

1.4 Objectives and Contribution

This study contributes to the expanding research between supervised and unsupervised learning approaches in intrusion detection system. Pre-processing and feature selection plays a crucial role in model performance, and ensemble learning methods can surpass independently working models by collaborating their robustness. Moreover, machine learning models provide scalable and resilient results in securing computerised networks against maturing cyber-attacks. The result of this study highlights the significance of combining multiple machine learning approaches to develop a strong intrusion detection system for it's implementation in real-world. The relevance of the existing study not only validates this research but also offers space for further enhancements in the stream.

2 Related Work

Intrusion Detection System plays a major role in securing networks against cyber-attacks. Conventional intrusion detection systems are functional with known attacks but usually are unable to detect rare attack patterns or to reduce false positives. With the advancements in the field of machine learning helped improving IDS abilities, utilizing supervised and unsupervised learning approaches, feature selection and ensemble models.

The proposed study extends the intrusion detection system related studies by utilizing both supervised and unsupervised learning models in an ensemble learning framework. The goal was to improve accuracy in detection, reduce false positives, and be scalable for real-time applications across computerised networks.

The author in "Machine Learning for Intrusion Detection Systems: A Systematic Overview" emphasizes feature selection and model evaluation metrics since this is just an overview of classification covering both supervised and unsupervised learning techniques(Stewart, Kolajo and Daramola, 2024). The current research adopts, feature selection module, utilizes SelectKBest, enabling the selection of those features most relevant to classification that contribute to both high model accuracy and efficiency. However, while assessing the

performance of ensemble model, evaluation metrics like accuracy, F1-score, and ROC-AUC are important(Stewart, Kolajo and Daramola, 2024).

The writer of "Intrusion Detection System Using Machine Learning Techniques: A Review" achieved the accuracy of >95% for both binary classification in the Random Forest and Gradient Boosting models (Musa *et al.*, 2020). Since Random Forest is robust and performs well in supervised learning tasks, it was included as the fundamental aspect of the current research. On the other hand, this study also addressed the issues in detecting unknown attack patterns by working with multi-class classification and combining Random Forest with anomaly detection models(Musa *et al.*, 2020).

The author of "Comparative Analysis of Intrusion Detection System Using Decision Trees" highlighted how decision trees are interpretable and computationally effective for small datasets(Azam, Islam and Huda, 2023). However, to maintain the interpretability along with improving accuracy for larger datasets, the current study leveraged Random Forest an ensemble of decision trees. By utilizing ensemble approach the research addresses the scalability issues which was identified in the previous research(Azam, Islam and Huda, 2023).

The author of "A Novel Time Efficient Approach of Smart Intrusion Detection System" focused mainly on lightweight models for intrusion detection in real time(Seth, Singh and Kaur Chahal, 2021). Regarding the author's focus on lightweight model for IDS in real time in the previous studies, in the current study, mainly focuses on efficiency using ensemble approach, considering computationally efficient anomaly detection methods such as the Isolation Forest and One-Class SVM. The current study address the limitation of the results being dataset specific, as this paper conducts model training on diverse data(Seth, Singh and Kaur Chahal, 2021).

The author of "Optimized Intrusion Detection Model for Identifying Known and Innovative Cyber Attacks Using SVM Algorithms" focuses on optimizing SVM models using feature selection techniques like PCA and Recursive Feature Elimination(Selvan, 2024). The current research also focuses on utilizing an ensemble learning technique- Voting Classifier which combines multiple machine learning models to enhance performance through feature selection techniques namely SelectKBest and RFE.

The author of "A Detailed Analysis of CIC-IDS 217 Dataset For Designing Intrusion Detection Systems" uses CIC-IDS 2017 dataset which is also utilized in the current research. However, the author (Panigrahi and Borah, 2024) focuses to clean an preprocess the dataset since it is highly imbalanced using feature selection, handling missing and imbalanced classes for further model training which exactly aligns with the current research.

The author proposed, base classifiers using Gaussian Naïve Bayes, Logistic Regression, and Decision Tree and integrated them using ensemble approach with Stochastic Gradient Descent to improve the detection performance and achieved high accuracy of 97.8%. (Thockchom, Singh and Nandi, 2023). However, this research is built on the foundation of the previous research leveraging ensemble approach which addressed the computational complexity by optimizing model selection and hyperparameters.(Thockchom, Singh and Nandi, 2023). To deal with the false positives problem, this study combines Logistic Regression, Random Forest with One-Class SVM, and Isolation Forest to improve robustness, accuracy in detecting and identifying unseen attack patterns, and overall reliability of the model.

The author of "Unsupervised and Ensemble Based Anomaly Detection for Network IDS" offered a practical example on how to utilize unsupervised and ensemble learning for network intrusion detection by combining multiple anomaly detection features, e.g. Mahalanobis distance and autoencoders to enhance anomaly detection. (Yang and Hwang, 2022). The current study incorporated ensemble framework to build a robust and reliable intrusion detection systems.

2.1 Quantitive Analysis

Sr.No.	Paper Title	Techniques Used	Dataset	Accuracy (%)	Strengths	Limitations
1	Machine Learning for Intrusion Detection Systems: A Systematic Overview(Stewart, Kolajo and Daramola, 2024)	Supervised Learning (SVM, Random Forest) Unsupervised (K-Means, Isolation Forest)	Overview	N/A	Offers vast classification of supervised and unsupervised machine learning techniques for intrusion detection systems.	Lacks implementation as it is just a theoretical research.
2	Intrusion Detection System Using Machine Learning Techniques: A Review(Musa *et al.*, 2020)	Random Forest, Gradient Boosting	CICIDS 2017	>95	Underlines the strengths of Random Forest and Gradient Boosting for Binary classification	Limited to binary classification and did not explore multiclass attacks or real time applicability.
3	Comparative Analysis of IDS Using Decision Trees(Azam, Islam and Huda, 2023)	Decision Trees	NSL-KDD	85-90	Decision trees are efficient for small datasets	Limited to one technique. Did not consider hybrid approach for performance enhancement.
4	A Novel Time Efficient Approach of	Light Gradient Boosting Machine	CICIDS 2018	97.73	Mainly works on balancing	Testing was done on the dataset

	Smart IDS (Seth, Singh and Kaur Chahal, 2021)				computational efficiency and detection accuracy.	which limited the general applicability.
5	A Detailed Analysis of CICIDS 2017 Dataset for Designing Intrusion Detection Systems. (Panigrahi and Borah, 2024)	Dataset Analysis, Class Imabalnce Handling, Feature Selection, Label for Balanced Data	CIC-IDS 2017 Dataset	Not directly shown in the paper	Offers a detailed analysis of the dataset. Highlights the limitations of the dataset like missing values.	Missing instances, class imbalance for certain attacks.
6	A Novel Ensemble Learning Model for IDS(Thockchom, Singh and Nandi, 2023)	Random Forest, Logistic Regression, Decision Tree (Stacking Ensemble)	NSL-KDD	97.8	The integration of multiple supervised models allows for performance with high precision.	High computation complexity and no strategy for reducing false positives.
7	Unsupervised and Ensemble Based Anomaly Detection for Network IDS(Yang and Hwang, 2022)	Isolation Forest, DBSCAN(Unsupervised Ensemble)	UNSW-NB15	90	Uses Unsupervised models for finding novel threats where no labelled data exists	A higher rate of false positives compared to supervised model.

Table 1: Quantitive Analysis

2.2 Research Gap and Justification

With all these developments, there are still some gaps in building a scalable, effective, and reliable intrusion detection systems which can manage different types of datasets, reduce false positives, and detect known and unknown threats both. In order to overcome these constraints, the current study combines anomaly detection- Isolation Forest, One-Class SVM, and supervised learning- Logistic Regression, Random forest in an ensemble model- Voting Classifier which is tailored for business networks. This study also also makes use of feature selection and performance evaluation using different types of evaluation metrics.

3 Research Methodology

The research methodology provides an idea about the structured plan followed to develop, evaluate and validate the intrusion detection system. This methodology is established in the previous researches and is developed upon well-known machine learning techniques and evaluation tactics.

The following steps were used while working on this research:

1. Dataset Selection
   The CIC-IDS 2017 Dataset was selected as it is popularly used in Intrusion Detection System research for the verity of attack patterns and practical network traffic behaviours.
   Justification: The chosen dataset comprises benign and malicious network traffic, making room for thorough assessment of supervised, unsupervised and ensemble models.

2. Data Preprocessing
   The preprocessing of the dataset was done to make sure the quality of data and model compatibility is exercised.
   a. Handling Infinite and Missing Values: Infinite values were replaced by zeros and mean column represents missing values.
   b. Label Encoding: 0 for benign and 1 for malicious were the numerical values which were encoded in the target variable "Label" using LabelEncoder.
   c. Feature Scaling: The make sure there was fair model comparison, features were standardized to mean of 0 and standard deviation of 1 using StandardScaler.

3. Feature Selection
   To reduce computation time and enhance model performance, choosing relevant features is important. Hence, based on ANOVA F-statistic, SelectKBest was used to recognize top 10 important features.

4. Model Development
   a. Supervised Models: The implementation of Logistic Regression for baseline comparison and Random Forest was done.
   b. Unsupervised Models: To detect novelty in the network traffic, Isolation Forest and One-Class SVM was used.
   c. Ensemble Approach: To enhance the accuracy and strength of the implemented model, an ensemble framework was also implemented by combining Logistic Regression, Random Forest Classifier, Isolation Forest, and One-Class SVM using Voting Classifier. Voting Classifier incorporates the strengths of both supervised and unsupervised models using soft voting which maximizes detection accuracy and minimizes false positives.

5. Execution Set-Up
   The execution set-up comprised of handling of dataset and splitting it into training and testing subsets, while implementing various exercises to evaluate the performance of the machine learning models to make sure the proposed intrusion detection system is robust enough.
   a. Dataset Splitting:
      The dataset was split into two subsets:
      - Training Set: 70% of the dataset was used to train the machine learning models. This subset helps the models to understand patterns and connections between features and target label, for instance, benign vs. malicious traffic.

- Testing Set: 30% of the dataset was reserved for assessing the performance of the trained models on unrevealed data for a neutral assessment of their capabilities.

Stratified sampling technique was used to maintain integrity of the dataset as it handles the original distribution of the target classes, e.g., benign, DDoS, etc. in training and testing sets both, as it keeps the underrepresented classes from being left out or lacking representation in the subsets, this is crucial for datasets with unequal class distribution and helps to avoid biased outputs.

3.1 Evaluation Methodology

Building on suggestions from the related studies, a thorough evaluation methodology was used to establish the effectiveness of the intrusion detection system(Tait *et al.*, 2021) (Ogundokun *et al.*, 2023).

1. Evaluation Metrics
   The evaluation of the models were done using the following metrics:
   - **Accuracy,** measures the adequacy of the model.
   - **Precision,** points out the portion of true positives from all positive prediction.
   - **Recall,** shows the model's potential to detect authentic positive es.
   - **F1-Score,** offering a stable performance metric, a concrete mean of precision and recall.
   - **ROC-AUC,** Receiver-operating characteristic curve Area under the curve: Evaluated the relationship between true positives and false positives rates, mainly for unbalanced dataset.
   - **Confusion Matrix,** a pictorial display of true positives, false positives, true negatives, false negatives.

2. Comparison with Existing Modern Approach
   The suggested intrusion detection system was compared step-by-step with the already existing studies to validate it's effectiveness. This comparison study depicts the potential advantage of the proposed framework over conventional models and benchmark it's effectiveness. Logistic Regression and Random Forest were the chosen baseline methods as they are well recognized for it's utilization in the field of machine learning based intrusion detection systems. Logistic Regression is computationally efficient model which is widely used and easy to understand. On the other hand, Random Forest is an ensemble model well-known to manage non-linear relation often used in IDS research. These baseline techniques focus on the advantages and disadvantages of the traditional approach while setting grounds for comparing modern approaches. Decision Tree, Gradient Boosting are some of the supervised learning models which are often mentioned in previous studies in intrusion detection systems for being effective in recognizing known attack patterns. While, unsupervised techniques like Isolation Forest and One-Class SVM have potential to identify novel attack types. However, in the recent times, studies have shown potential of hybrid models and how combing both supervised and unsupervised learning models can enhance detection accuracy and decrease false positives.

3. Performance Results
   The results showed that, however logistic regression offered adequate results for binary classification, it did not work well with multiclass frameworks and detecting

complicated attack patterns. While on the other hand, Random Forest's performance was superior and but had some constraints in managing novel attacks. Although Isolation Forest and One-Class SVM worked exceptionally well in identifying anomalies, they struggled with high amount of false positive results. Nonetheless, an ensemble approach results showed superior performance by combining the strengths of supervised and unsupervised models with respect to the accuracy, F1-score, and ROC-AUC along with successfully reducing the false positives.

4. Key Findings
   The development of ensemble-based IDS accomplishes notable improvement over baseline models in predicting and detecting known and unknown attack patterns. It also reduced false positives which addressed the anomaly detection struggle. By combining supervised and unsupervised learning approaches, it provides intrusion detection system to perform with verity of attack types and also improves it's dependability.

5. Justification for the Approach Used
   By demonstrating how well the suggested intrusion detection system can overcome the drawbacks of conventional and modern models both, this comparison justifies the research approach of the system. Comparing the systems to recognized approaches displays that ensemble model provides a reliable, scalable and accurate way to secure corporate networks in real time.

3.2 Statistical Analysis

To understand the class distribution, feature relationships, and any possible data imbalances, exploratory data analysis approach was used to study the dataset. ANOVA F-statistic was used to evaluate the statistical significance of the features selected. Performance was evaluated with the obtained results and comparing the evaluation scores.

4 Design Specification

This system enhanced Intrusion Detection System by utilizing machine learning techniques to categorize network traffic data as benign or malicious.

4.1 System Functionality

1. The first stage of the system was Data Preprocessing which involves cleaning and preprocessing the dataset which contains network traffic data.
   - Handling Missing Values: The missing values in the dataset were handled using SimpleImputer which handles the missing values of the dataset by substituting the mean of the corresponding column.
   - Feature Scaling: Normalization of the features is done using StandardScaler to make sure that all numeric features provide equally to learning models, preventing any bias because of varying size.
   - Label Encoding: LabelEncoder was employed to encode the benign or malicious network traffic into binary format which was under the categorical column "Label".

- Feature Selection: To select the top 10 most relevant features from the dataset, SelectKBest with ANOVA F-statistic "f_classif" which is a feature selection technique was used.

2. Multiple machine learning models were trained and assessed to classify the network traffic as benign or malicious. The following are the models chosen which can handle both classification and anomaly detection.
    - Logistic Regression: This classifier is employed as the baseline model for binary classification. This model evaluates the relation among features to calculate the possibility of the classes by using logistic function. The logistic function maps input features to a possibility of binary value (0 and 1) underlining the probability of the input belonging to a specific class.
    - Random Forest Classifier: Here, Random Forest Classifier was used due to its capability of managing complex and non-linear relations between the features. Independent decision are made with the use of ensemble of decision trees which the model comprises of that are trained on a random subset of the data which gives the final prediction by gaining majority votes across all the trees. This helps in minimizing the overfitting and offers robust classification results mainly while working with complex datasets.
    - Isolation Forest: Instead of isolating normal data, isolation forest isolates the anomalies, mainly because it is easier to isolate the anomalies compared to normal data as they stay away from the majority of data.. Since the model is built on a series of decision trees, the outliers are isolated quicker compared to normal cases. This mechanism predicts anomalies by calculating how fast they are isolated in the trees.
    - One-Class SVM: This is the other anomaly detection approach which classifies major chunk of class as hyperplane and detects anything outside this area as anomaly. This approach works best where there are less or no labelled anomalies in the dataset. The kernel used in this approach was radial basis function which allows the model to catch complex and non-linear relations.
3. An Ensemble model was implemented to improve the performance and reliability of the model by using Voting Classifier approach. What it does is, it combines the predictions from Logistic Regression, Random Forest, Isolation Forest , and One-Class SVM, applies soft voting which averages the predictions of these models and gives the output. This technique reduces the errors of individual models and enhances generalization on unseen data. It also helps to reduce overfitting by integrating multiple models.

4.2 Framework and Libraries Used

This study relies on multiple libraries for data preprocessing and machine learning. Pandas was imported to read, manipulate, handle missing values, feature extraction and data transformation. NumPy offers support for numerical operations, array manipulation and mainly during the scaling and feature selection. Scikit-learn is a library for machine learning which comprises implementation of logistic regression, random forest, one-class SVM, voting classifier, while offering tools for feature selection, and model evaluation metrics, and confusion matrices. However, the implementation of this study did not include deep learning approach, even then TensorFlow was employed for the possibility of model extension in the future. Matplotlib and Seaborn was utilized for visualization of the model performances like plotting the ROC-AUC curve, etc.

4.3 System Architecture

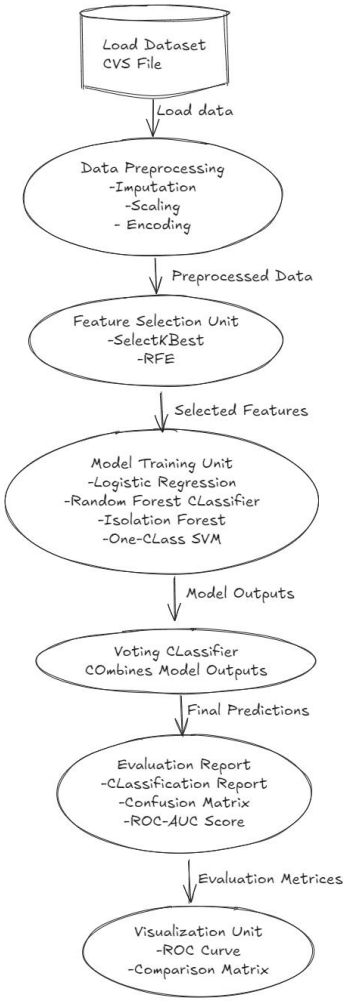

Image 1: System Architecture

This system is designed with the capability of handling different types of network traffic data and various machine learning and anomaly detection techniques for accurate classification, hence providing a very flexible framework of network intrusion detection.

5 Implementation

The final stage of the implementation of Intrusion Detection System, it includes implementing multiple machine learning models along with ensemble learning approach using Voting Classifier. The main objective of this implementation is to detect and classify network data more precisely as malicious or benign, reducing false positives in detecting complex attack patterns.

5.1 Data Preprocessing and Transformation

The CSV file of the network traffic dataset was used which comprised of multiple features like, packet lengths, flow duration, and many other different types of network columns. In the preprocessing stage the dataset was cleaned by first handling the missing values, scaling the numerical feature, and encoding categorical labels. Missing values were imputed by mean and

the NaN values were replaced by zeros. Feature scaling was done using StandardScaler to make sure there is consistency and LabelEncoder was used to encode the class labels to binary. This step made sure that the data was clean and usable for model training.

```
Selected features: Index(['Destination Port', 'Bwd Packet Length Max', 'Bwd Packet Length Mean',
       'Bwd Packet Length Std', 'Min Packet Length', 'Packet Length Mean',
       'Packet Length Std', 'URG Flag Count', 'Average Packet Size',
       'Avg Bwd Segment Size'],
      dtype='object')
```

Image 2: Feature Selection after Preprocessing

5.2 Model Development

Multiple machine learning models namely Logistic Regression as baseline for binary classification, Random Forest Classifier to train multiple decision trees and collect the prediction, Isolation Forest, and One-Class SVM were used to detect anomalies that are unseen traffic patterns, for developing this system.
Model training was done using the pre-processed data and it's performance was evaluated by using evaluation metrices like precision, recall, F1-score, confusion matrix and ROC_AUC curve.

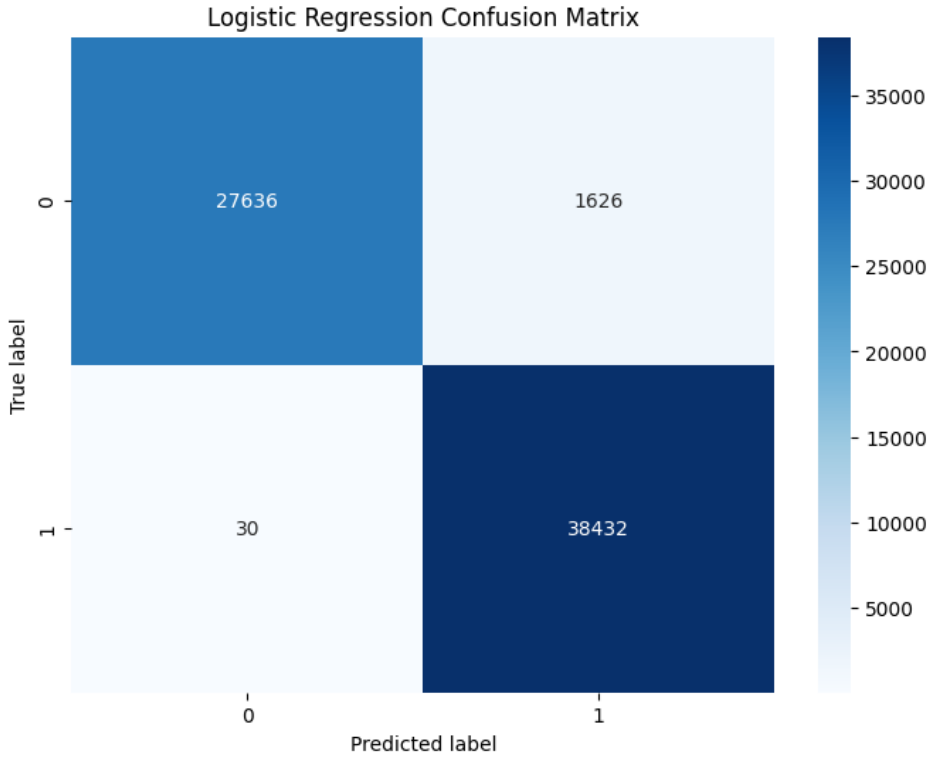

Image 3:Confusion Matrix- Logistic Regression

The Logistic Regression model's accuracy score is 94% and performs as required. Class 1 which is malicious traffic, is detected with accurate recall of 1.00. This says that the model perfectly identifies all the malicious traffic. The model is precise for class 0 which is benign but there's a slight struggle with recall being 0.86 because of moderate amount of false positives. However, the overall model is functional and can differentiate between benign and malicious traffic shown by ROC-AUC curve score 0.93.

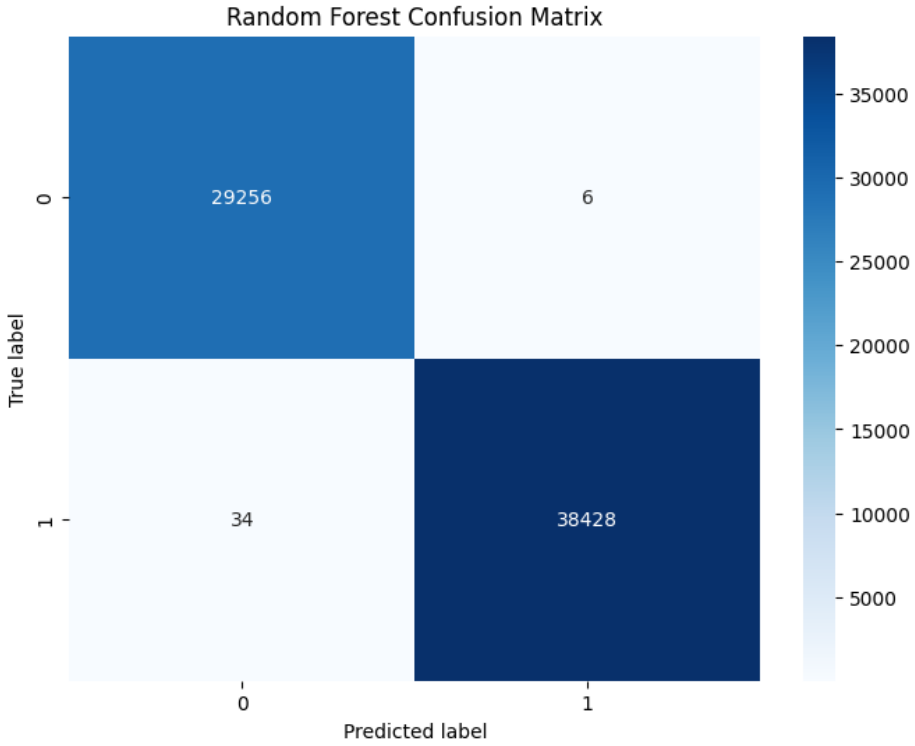

Image 4: Confusion Matrix- Random Forest Classifier

The performance of Random Forest was perfect with precision, recall and F1-score for both benign and malicious classes. The accuracy was 1.00 which states that the model made no wrong predictions. The ROC-AUC was 1.00 which shows that the model could differentiate between benign and malicious traffic both. The confusion matrix showed that the false positives were only 7 and false negatives were 30 which is very low. Random Forest worked extremely well on this dataset which says that this model is ideal for classification.

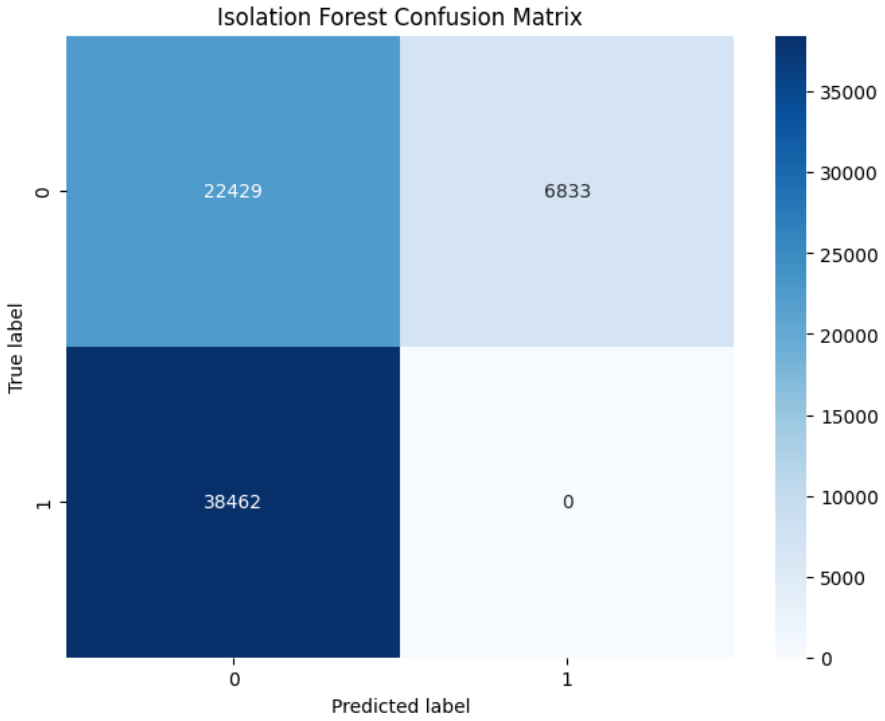

Table 5: Confusion Matrix breakdown- Isolation Forest

Isolation Forest model on its own was not suitable for this evaluation with accuracy of 0.33. It struggled in detecting the malicious traffic which is normal in anomaly detection models. Recall for class 1 was 0 which means the model missed all the malicious traffic, moreover the precision for benign traffic was 0.37 which is low too highlighting that there are many false positives. F1-score for benign was 0.50 and malicious was 0.00 which indicates the model's detection ability. The confusion matrix shows that the model could only predict benign traffic till a certain extent but could not identify malicious traffic which gave a lot of false positives. The ROC-AUC score was 0.38 which confirmed the model's capability to differentiate between benign and malicious traffic.

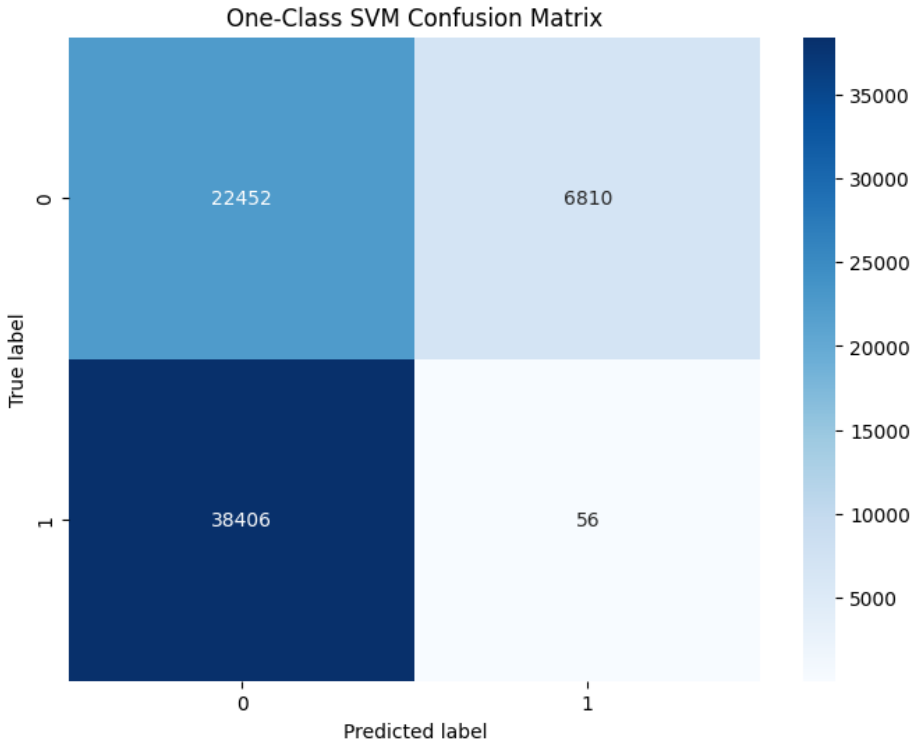

Image 6: Confusion Matrix-One-Class SVM

Precision for the malicious traffic is very low being 0.01 which says that the model predicted almost all the traffic as benign. For malicious traffic the recall was 0.00 which means the model failed to identify the malicious traffic. The accuracy score is 0.33 which is low too indicates the model's bad performance. In the confusion matrix it shows the model failed to detect malicious traffic as 38406 instances were misclassified as benign. The model also gave high false positives for benign traffic as 6810 were classified as malicious traffic. The ROC-AUC score of 0.38 indicates the efficiency of the model is low and it cannot distinguish between benign and malicious.

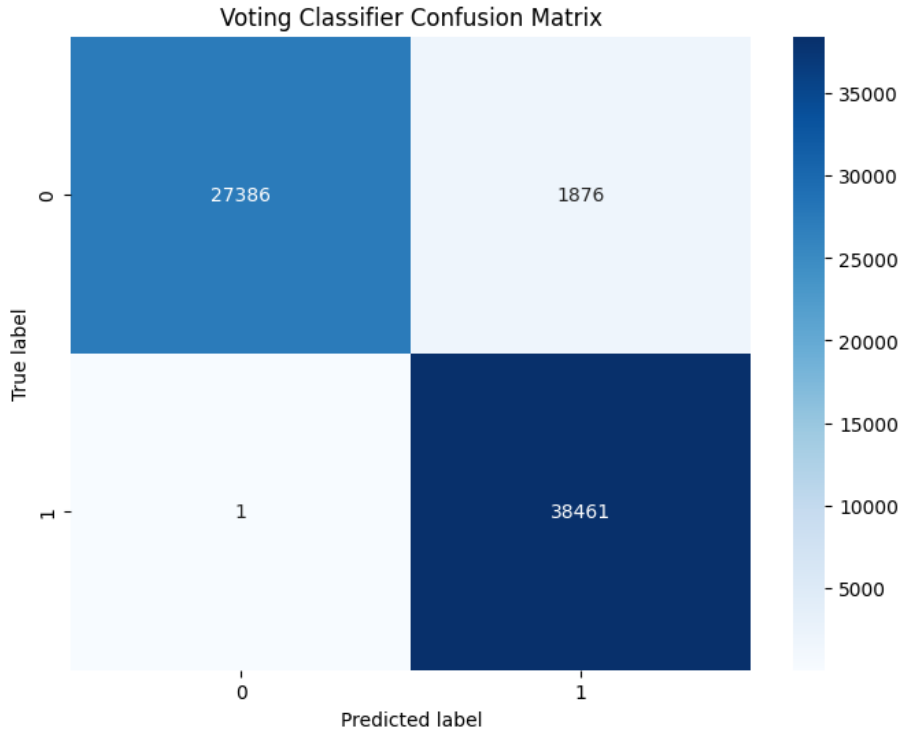

Image 7:Confusion Matrix- Voting Classifier

Voting Classifier performs outstandingly in this case, with high accuracy 0.97 along with excellent recall for malicious traffic which was 1.00. This highlights the model's reliability in detecting malicious traffic. Precision for benign was 1.00 which underlines the model's ability to avoid false positives. The model shows a tiny lower precision for malicious traffic with 0.96 which indicates that the model does misclassify some malicious traffic as benign. The ROC-AUC score of 0.97 which confirms the models reliability to differentiate between benign and malicious network traffic.
The Voting Classifier is extremely effectively in this study, shows exceptional performance in detecting malicious network traffic and avoiding false positives both.

5.3 Model Evaluation

The evaluation of all the model performance was done using multiple evaluation metrics to assess their capability to analyse the network traffic. Metrices such as;

Accuracy shows the correctness of the model. The accuracy is the ratio of rightly predicted instances to total instances. (Kumar, 2020)

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Total\ Instances}$$

Accuracy score offers an overall working of the intrusion detection system and calculates the amount of true and false predictions with respect to total number of instances. For this study, the accuracy score helped to understand how the model differentiates between benign and malicious network traffic. High accuracy means that the model is making right predictions. For example, Accuracy = (TP+TN)/(P+N) where P and N are total Positive and negative classes. (2+7)/(3+7)= 0.9 which means that the model is able to predict rightly 90% of the times.

Precision is the ratio of actual positives out of all the positive predictions the model actually made. (Kumar, 2020)

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

Precision calculates how well the model can find malicious traffic in relation to all cases where it predicted malicious traffic in this project, that is, building an IDS model to detect different kinds of cyber-attacks such as Ddos. True Positives define the number of real attacks which were correctly classifies as malicious in this model. False positives is the number of valid traffic incidents that the IDS model wrongly classified as malicious.

Recall is the actual number of malicious traffic which means the correct number of true positives.(Kumar, 2020)

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

Recall measures how successfully the model detects all malicious traffic, true positives in this project. The IDS detects different kinds of cyber attacks while false negatives indicate the missed ones by the model. High recall rate means that the system is effective and can prevent intrusions.

F1-Score is when precision and recall balance out mainly in unbalanced datasets.(Kumar, 2020)

$$F1 - Score = \frac{2\ x\ \ Precision\ x\ Recall}{Precision + Recall}$$

The F1-score is an important metric for evaluating IDS as it considers both precision and recall. It allows to evaluate how well the system maintains balance between attack detection and false positive reduction. The higher the F1-score the better the detection of malicious activities by an IDS.

ROC-AUC is the relation of true positive rates against false positive rate was assessed by considering the ROC curve to present the performance of the proposed model. It evaluates the effectiveness of the model, in this study the ROC-AUC score of Logistic Regression and Random forest were 0.97 and 1.00 respectively which states that the model could differentiate between benign and malicious traffic.(Kumar, 2020)

Confusion matrix helps in explaining the model mistake by proper description through true positives, false positives, true negatives, and false negatives.(Kumar, 2020)

6 Evaluation

This study was to enhance the Intrusion Detection System by using supervised learning approaches Logistic Regression, Random Forest, unsupervised learning Isolation Forest, One-Class SVM, and Voting Classifier ensemble model. The evaluation focuses on the reliability of these models to categorize network traffic as normal or malicious by reducing false positives and increasing detection accuracy.

6.1 Model Performance Analysis

6.1.1 Logistic Regression Evaluation

```
Logistic Regression Evaluation:
              precision    recall  f1-score   support

           0       1.00      0.94      0.97     29262
           1       0.96      1.00      0.98     38462

    accuracy                           0.98     67724
   macro avg       0.98      0.97      0.97     67724
weighted avg       0.98      0.98      0.98     67724

[[27636  1626]
 [   30 38432]]
ROC-AUC: 0.97
```

Image 8: Logistic Regression Evaluation

Accuracy: 98%
Precision: 0.96 for malicious traffic
Recall: 1.00 for malicious traffic
F1-Score: 0.98
ROC-AUC Score: 0.97
Logistic Regression model accomplished amazing recall results for detecting malicious network traffic which shows that it was capable of rightfully identify all the malicious instances. Although, the precision rate is slightly low to 0.96 for malicious traffic which shows that there were few false positives. However, the model's accuracy was 98% and its F1-score was 0.98 which is pretty strong but it also indicated certain limitations in handling complex and multi-class patterns.

6.1.2 Random Forest Evaluation

```
Random Forest Evaluation:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     29262
           1       1.00      1.00      1.00     38462

    accuracy                           1.00     67724
   macro avg       1.00      1.00      1.00     67724
weighted avg       1.00      1.00      1.00     67724

[[29256     6]
 [   34 38428]]
ROC-AUC: 1.00
```

Image 9: Random Forest Evaluation

Accuracy: 100%
Precision: 1.00 for both benign and malicious traffic

Recall: 1.00 for both benign and malicious traffic
F1-Score: 1.00
ROC-AUC: 1.00
Random Forest performed perfectly in all the metrices. It could accomplished a perfect accuracy, precision, recall, and F1-score of 1.00. The results shows that the model was able to accurately classify benign and malicious traffic. However, even though the results were excellent it couldn't detect novel attacks.

6.1.3 Isolation Forest (Anomaly Detection)

```
Anomaly Detection (Isolation Forest) Evaluation:
              precision    recall  f1-score   support

           0       0.37      0.77      0.50     29262
           1       0.00      0.00      0.00     38462

    accuracy                           0.33     67724
   macro avg       0.18      0.38      0.25     67724
weighted avg       0.16      0.33      0.22     67724

[[22444  6818]
 [38462     0]]
ROC-AUC: 0.38
```

Image 10: Isolation Forest Evaluation

Accuracy: 33%
Precision: 0.00 for malicious traffic
Recall: 0.00 for malicious traffic
F1-Score: 0.00
ROC-AUC Score: 0.38
Isolation Forest did not perform well with the accuracy score of 33%. The precision, recall and F1-score was malicious traffic was 0 which indicates that the model failed to detect malicious traffic. The ROC-AUC score of 0.38 which showed that the model was incapable of distinguishing malicious and benign network traffic.

6.1.4 One-Class SVM (Anomaly Detection)

```
Anomaly Detection (One-Class SVM) Evaluation:
              precision    recall  f1-score   support

           0       0.37      0.77      0.50     29262
           1       0.01      0.00      0.00     38462

    accuracy                           0.33     67724
   macro avg       0.19      0.38      0.25     67724
weighted avg       0.16      0.33      0.22     67724

[[22452  6810]
 [38406    56]]
ROC-AUC: 0.38
```

Image 11: One-Class SVM Evaluation

Accuracy: 33%
Precision: 0.01 for malicious traffic
Recall: 0.00 for malicious traffic
F1-Score: 0.00
ROC-AUC: 0.38
Just like Isolation Forest, One-Class SVM performed extremely poor. The accuracy it accomplished was 33% and the precision for detecting malicious network traffic was 0.01 which indicates that it did not detect any true positives. The F1-score and recall value was 0.00 too which stated that the model failed to detect any malicious traffic. ROC-AUC score was 0.38 which was also low and showed that the model struggled to detect anomalies.

6.1.5 Voting Classifier (Ensemble Model)

```
Voting Classifier (Combined Model) Evaluation:
              precision    recall  f1-score   support

           0       1.00      0.94      0.97     29262
           1       0.96      1.00      0.98     38462

    accuracy                           0.97     67724
   macro avg       0.98      0.97      0.97     67724
weighted avg       0.98      0.97      0.97     67724

[[27587  1675]
 [   19 38443]]
ROC-AUC: 0.97
```

Image 12: Voting Classifier Evaluation

Accuracy: 97%
Precision 0.96 for malicious traffic
Recall: 1.00 for malicious traffic
F1-Score: 0.98
ROC-AUC score: 0.97
Voting Classifier used soft voting which picks up the prediction from each model. Every classifier allocates a prediction to each class, and the ensemble model's prediction is the class with the highest probability (Ahmed, 2023). It is the combination of Logistic Regression, Random Forest, Isolation Forest, and One-Class SVM, performed extremely well. An ensemble model accomplished the 1.00 as its recall value for malicious network traffic which indicates that the model was successful to detect all the malicious instances. The precision and F1-score was 0.96 and 0.98 respectively, which is solid. The ROC-AUC score was 0.97 which was as good as Logistic Regression when compared but with much better robustness as multiple models were combined.
The evaluation of the models showed that the Voting Classifier notably improved the system's ability to decrease false positives and to detect anomalies when compared with the performance of individual models.

6.2 Visualization of ROC Curves

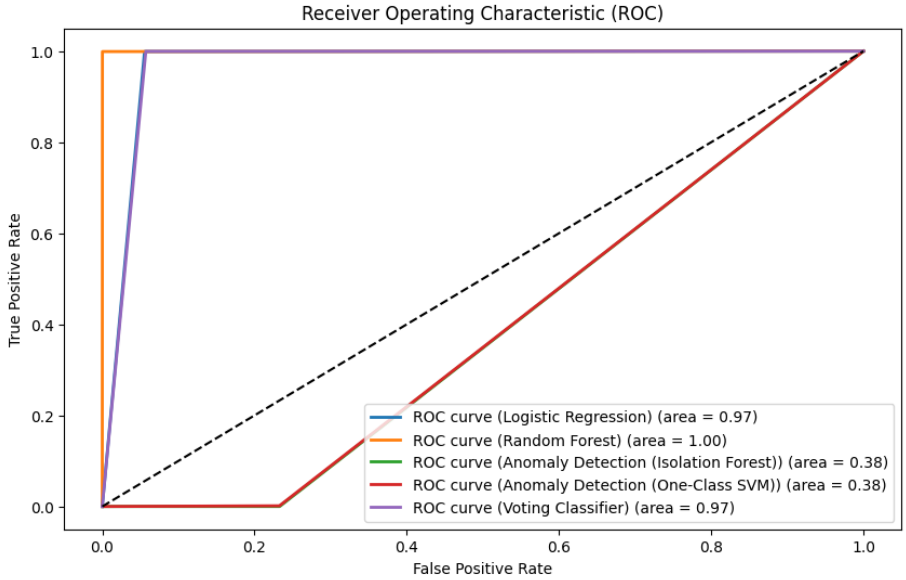

Image 13: ROC Curve

ROC Curve of all the models was plotted for better visualization of the performances. This curve plots true positive rate against false positive rate, and AUC is the model's ability to differentiate between malicious and normal network traffic.

In the above plotting, Logistic Regression and Random Forest are showing high AUC rate which is 0.97 and 1.00 respectively pointing outstanding performance in differentiating. the anomaly detection models, Isolation Forest and One-Class SVM AUC rates are very low indicating poor performance of the models. However, The AUC of Voting Classifier is 0.97 demonstrating powerful performance of the overall model.

6.3 Visualization of Comparison Matrix

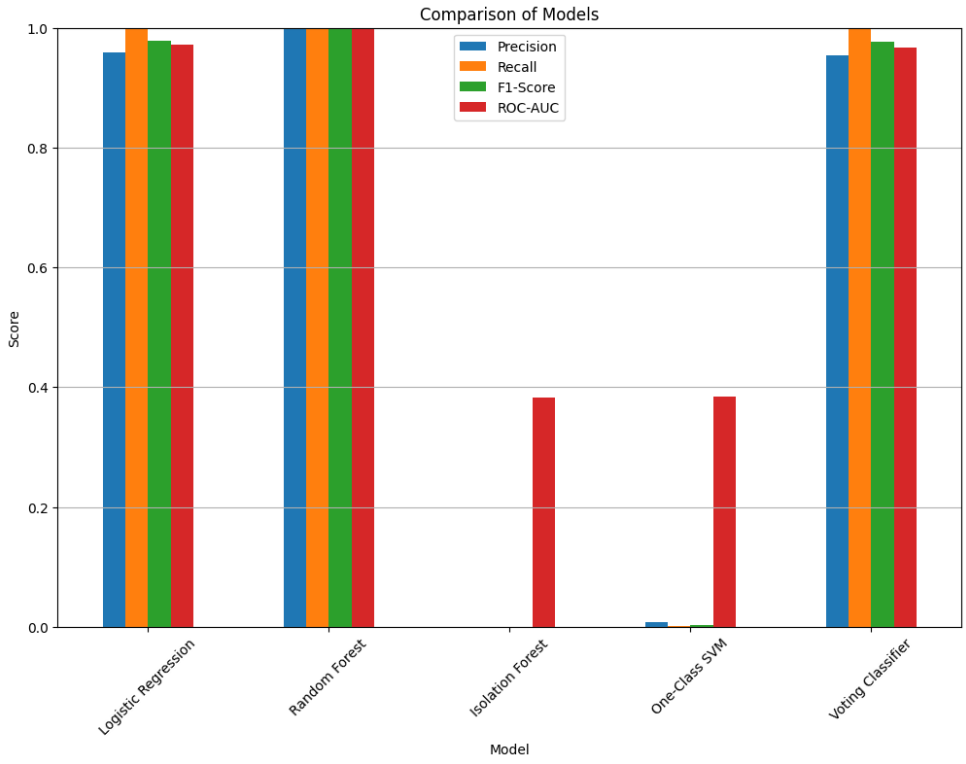

Image 14: Comparison of Models Graph

The graph above compares models using four important metrices: Precision, Recall, F1-Score, and ROC-AUC score. The Voting Classifier enhances overall performance by taking advantage of the multiple models. It performs best in every evaluation parameter, specially in terms of reducing false positives and detecting both seen and unseen attack patterns. However, the presence of imbalanced datasets creates problems for both Isolation Forest and One-Class SVM, leading to poor performance. While these models are valuable in anomaly detection, they are not suitable where minimization of false positives and false negatives is most important. While Random Forest and Logistic Regression performs relatively well on this dataset, Random Forest works perfectly in every aspect, specially for ROC-AUC and F1-score. According to the evaluation of each model, the best method for intrusion detection system is an ensemble learning- Voting Classifier as it notably outperforms each of the individual models.

6.4 Statistical Analysis

	Model	Accuracy	Precision	Recall	F1-Score
0	Logistic Regression	0.971827	0.959409	0.999220	0.978910
1	Random Forest	0.999455	0.999844	0.999116	0.999480
2	Isolation Forest	0.383501	0.000000	0.000000	0.000000
3	One-Class SVM	0.384365	0.008156	0.001456	0.002471
4	Voting Classifier	0.971132	0.958248	0.999506	0.978442

Image 15: Comparison Matrix

The models where compared based on the evaluation matrices like classification report, comparison matrix, and ROC-AUC score. The comparison of models was done using there performance matrix to assess what model worked the best. Image 9 shows the model performance with it's accuracy, precision, recall, and F1-score respectively.

6.5 Discussion

The main objective of this study was to improve the performance of Intrusion Detection System using Machine Learning approaches to detect various network intrusions like, DDoS, Brute Force attack, SQL Injections, etc. In this research, the efficiency of several machine learning models for IDS was reviewed, such as ensemble model-Voting Classifier, Random Forest, Isolation Forest, One-Class SVM, and Logistic Regression. Among these, Random Forest model performed best, with perfect score in the ROC-AUC score of 1.00, which proved its capacity to manage complex, non-linear communication in the network. Logistic Regression, on the other hand, struggles with complicated attack patterns, with respect to the ROC-AUC score of 0.97. The scores by Isolation Forest and One-Class SVM were 0.38 insinuating their failure at finding malicious traffic and making many false positives.(Chandola, no date) This showed the weakness of these models when considering the dataset was largely imbalanced, with benign traffic being mostly greater than malicious traffic, although each is strong in identifying anomalies. The Voting Classifier combined the predictions of all the models and yielded an ROC-AUC score of 0.97, a bit lower than that of Random Forest but considerably higher than the anomaly detection models. This again underscores the strength of ensemble learning, whereby weak models can show balanced detection accuracy to reduce false positives when combined. (Meryem and Ouahidi, 2020)
However, the limitation in the design of experiments include problems with data imbalance

and class distribution that generated false positives in anomaly detection models. The study academically consolidates the efficiency of ensemble learning methods in IDS and the importance of data preprocessing. Practically, it shows that organizations can enhance IDS by leveraging ensemble models and giving attention to class imbalance and feature engineering for better detection accuracy.

7 Conclusion and Future Work

Network traffic should be secure for smooth operational function of an organization. An intrusion detection system works in protecting the organization from any malicious intrusions. The previous studies have used ensemble approach but the combination of supervised and unsupervised learning method combined with Voting Classifier for anomaly detection was not utilized. This study has presented the fact that anomaly detection models increase the efficiency of intrusion detection systems by 97% when combined with ensemble learning models. Specifically, it was established that Voting Classifier holds great potential to increase the detection accuracy by reducing false positives and delivering highly reliable IDS solution. Future research can focus on solving problems related to data imbalance using strategies like SMOTE, class weighing or under sampling, will greatly improve performance. The scalability can also be maximized by hybrid models, where sophisticated classifiers are combined with the computationally efficient ones. This could further be implemented in the future using techniques like cross-validation at 10 fold would be much more robust and results in less overfitting. Feature selection could be improved further using RFE. Also, model selection can be optimized for speed and scalability using computationally efficient models like Logistic Regression together with highly accurate models such as Random Forest. Future studies can also make real-time attack detection in IDS more effective by enhancing the computational speed and resource efficiency like enhanced dataset for an ensemble model, thereby making them suitable for commercial use in cybersecurity applications.

References

Ahmed, I. (2023) ‘What is Hard and Soft Voting in Machine Learning?’, *Medium*, 31 May. Available at: https://ilyasbinsalih.medium.com/what-is-hard-and-soft-voting-in-machine-learning-2652676b6a32 (Accessed: 12 December 2024).

Azam, Z., Islam, Md.M. and Huda, M.N. (2023) ‘Comparative Analysis of Intrusion Detection Systems and Machine Learning-Based Model Analysis Through Decision Tree’, *IEEE Access*, 11, pp. 80348–80391. Available at: https://doi.org/10.1109/ACCESS.2023.3296444.

Disha, R.A. and Waheed, S. (2022) ‘Performance analysis of machine learning models for intrusion detection system using Gini Impurity-based Weighted Random Forest (GIWRF) feature selection technique’, *Cybersecurity*, 5(1), p. 1. Available at: https://doi.org/10.1186/s42400-021-00103-8.

gmcdouga (2024) *Check Point Research Reports Highest Increase of Global Cyber Attacks seen in last two years – a 30% Increase in Q2 2024 Global Cyber Attacks, Check Point Blog*. Available at: https://blog.checkpoint.com/research/check-point-research-reports-highest-increase-of-global-cyber-attacks-seen-in-last-two-years-a-30-increase-in-q2-2024-global-cyber-attacks/ (Accessed: 10 December 2024).

IDS 2017 | Datasets | Research | Canadian Institute for Cybersecurity | UNB (no date). Available at: https://www.unb.ca/cic/datasets/ids-2017.html (Accessed: 9 December 2024).

Kumar, S. (2020) ‘8 Important Evaluation Metrics for Classification Models’, *AITUDE*, 26 March. Available at: https://www.aitude.com/8-important-evaluation-metrics-for-classification-models/ (Accessed: 6 December 2024).

Musa, U.S. *et al.* (2020) ‘Intrusion Detection System using Machine Learning Techniques: A Review’, in *2020 International Conference on Smart Electronics and Communication (ICOSEC). 2020 International Conference on Smart Electronics and Communication (ICOSEC)*, pp. 149–155. Available at: https://doi.org/10.1109/ICOSEC49089.2020.9215333.

Narkhede, S. (2021) *Understanding Confusion Matrix, Medium*. Available at: https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62 (Accessed: 11 December 2024).

Ogundokun, R.O. *et al.* (2023) ‘Intrusion Detection Systems Based on Machine Learning Approaches: A Systematic Review’, in *2023 International Conference on Science, Engineering and Business for Sustainable Development Goals (SEB-SDG). 2023 International Conference on Science, Engineering and Business for Sustainable Development Goals (SEB-SDG)*, pp. 01–04. Available at: https://doi.org/10.1109/SEB-SDG57117.2023.10124506.

Panigrahi, R. and Borah, S. (2024) ‘(PDF) A detailed analysis of CICIDS2017 dataset for designing Intrusion Detection Systems’, *ResearchGate* [Preprint]. Available at: https://www.researchgate.net/publication/329045441_A_detailed_analysis_of_CICIDS2017_dataset_for_designing_Intrusion_Detection_Systems (Accessed: 11 December 2024).

Priyanka (2022) ‘Beyond Accuracy: Recall, Precision, F1-Score, ROC-AUC’, *Medium*, 6 November. Available at: https://medium.com/@priyankads/beyond-accuracy-recall-precision-f1-score-roc-auc-6ef2ce097966 (Accessed: 11 December 2024).

Selvan, M.A. (2024) ‘Svm-Enhanced Intrusion Detection System for Effective Cyber Attack Identification and Mitigation (1st edition)’, *Journal of Science Technology and Research (JSTAR)*, 5(1), pp. 397–403.

Seth, S., Singh, G. and Kaur Chahal, K. (2021) ‘A novel time efficient learning-based approach for smart intrusion detection system’, *Journal of Big Data*, 8(1), p. 111. Available at: https://doi.org/10.1186/s40537-021-00498-8.

Stewart, D., Kolajo, T. and Daramola, O. (2024) ‘Machine Learning for Intrusion Detection Systems: A Systematic Literature Review’, in K. Arai (ed.) *Proceedings of the Future Technologies Conference (FTC) 2024, Volume 1*. Cham: Springer Nature Switzerland, pp. 623–638. Available at: https://doi.org/10.1007/978-3-031-73110-5_42.

Tait, K.-A. *et al.* (2021) ‘Intrusion Detection using Machine Learning Techniques: An Experimental Comparison’. arXiv. Available at: https://doi.org/10.48550/arXiv.2105.13435.

Thockchom, N., Singh, M.M. and Nandi, U. (2023) ‘A novel ensemble learning-based model for network intrusion detection’, *Complex & Intelligent Systems*, 9(5), pp. 5693–5714. Available at: https://doi.org/10.1007/s40747-023-01013-7.

Yang, D. and Hwang, M. (2022) ‘Unsupervised and Ensemble-based Anomaly Detection Method for Network Security’, in *2022 14th International Conference on Knowledge and Smart Technology (KST)*. *2022 14th International Conference on Knowledge and Smart Technology (KST)*, pp. 75–79. Available at: https://doi.org/10.1109/KST53302.2022.9729061.