INTRUSION DETECTION SYSTEM USING MACHINE LEARNING

Penumati Harish Chandra
2205576
Dept. of Computer Science
University of Houston
Houston, TX - USA
hpenumat@cougarnet.uh.edu

Pavuluri Yaswanth Chowdary
2183774
Dept. of Computer Science
University of Houston
Houston, TX - USA
ypavulur@cougarnet.uh.edu

Senagapalli Sai Praneeth
2199053
Dept. of Computer Science
University of Houston
Houston, TX - USA
ssenigap@cougarnet.uh.edu

I. ABSTRACT

***Abstract*:** Intrusion Detection Systems (IDS) is a networking classification task in which we detect, classify, and prevent illicit network traffic. This research explores the development and implementation of machine learning models for intrusion detection systems (IDS). We propose a novel hybrid model that combines the strengths of Logistic Regression and Decision Trees to achieve superior performance in classifying benign and malicious network traffic. The proposed model outperforms standalone models in terms of accuracy, precision, recall, F1 score, and ROC AUC score. The findings suggest that hybrid models offer a promising approach for enhancing the effectiveness of IDS.

Keywords: - Intrusion Detection systems, Hybrid Model, Logistic Regression, Decision Trees, and ROC AUC score.

II. LITERATURE SURVEY

Vinayakumar et al, **[1]** introduced a deep learning based model for intrusion detection systems based on Recurrent Deep learning. Identifying optimal features of upcoming traffic for the machine learning model is one of the key challenges. The model proposed first do the kernel-based principal component (KPCA) analysis and then use these feature in recurrent models for the classification task. The described model might lack robustness and generalizability when exposed to an adversarial modified dataset.

Tanzila et al., **[2]** obtainable an Anomaly-based intrusion Detection System for IoT Networks through a Deep Learning Model. This study focus on intrusion Detection System especially for IOT devices which is a critical challenge. A Machine Learning Model can accurately classify the traffic into normal and malicious. A CNN-based model has the ability to deal with whole traffic in terms of features. The model underwent training and evaluation using the NID Dataset, achieving an accuracy of 99.51%.

The authors **[3]** presented a deep learning-based model for countering the escalating threat of cyberattacks on fully integrated servers, applications, and communication networks in the situation of the Internet of Things (IoT). The use of an autoencoder model particularly stood out, exhibiting superior performance by significantly reducing detection time and enhancing precision.

Logeswari et al. presented **[4]** a cutting-edge solution to address the security challenges posed by Software Defined Networking (SDN). To tackle these security concerns, the authors propose a novel Hybrid Feature Selection algorithm coupled with the ***LightGBM Intrusion Detection System*** (HFS-

LGBM IDS). The methodology unfolds in two distinct phases. In the starting step, the Correlation-based Feature Selection (CFS) algorithm is employed to reduce data dimensions and derive an optimal feature subset. The Random Forest Recursive Feature Elimination (RF-RFE) is utilized in the second phase to refine and finalize the feature set then leverage the LightGBM algorithm for the detection and classification of various types of network attacks. The experiment accompanied on the ***NSL-KDD*** dataset demonstrates the efficacy of the HFS-LGBM IDS.

Dhia et al., published research **[5]** on a deep learning-based model for enhancing the security of Internet of Things (IoT) devices through the implementation of an improved Intrusion Detection System (IDS). The complexities of intrusion detection in the IoT environment, the authors focused on the crucial role of feature extraction algorithms in shaping the detection accuracy. The results showcased the effectiveness of the VGG-16 model in conjunction with stacking, yielding an impressive accuracy of 98.3%.

Farooq et al., **[6]** provided a research on Machine Learning Approach for an Intrusion Detection System with fusion. The projected IDS-FMLT system model achieved a validation accuracy of 95.18% and a low miss rate of 4.82% in intrusion detection.

In the paper **[7]**, the emphasis is on addressing the weakness of machine learning-based intrusion detection systems (IDSs) to adversarial attacks within the realm of network security. The primary objective of the paper is to reduce the susceptibility of machine learning classification models employed in IDSs to adversarial attacks. The proposed framework consists of two pivotal phases: initialization and detection. During the initialization phase, the authors utilize a support vector machine classification model to train the IDS. In the second phase, the analysis results of the classification performed by the trained IDS are examined using the extracted features. To evaluate the effectiveness of the proposed method, the researchers employed the NSL-KDD dataset, a widely recognized benchmark in the field of intrusion detection.

III. INTRODUCTION

An essential part of cybersecurity are intrusion detection systems (IDS), which protect networks from hostile activity and illegal access. As sentinels, they keep an eye on network traffic, looking for irregularities and unusual patterns that might point to an attempted incursion. IDS can identify and notify security staff of possible threats by examining system calls, network packets, and traffic logs. Host-based and network-based IDS are the two primary types available. Host-based intrusion detection systems (IDS) are installed on specific PCs or servers and keep an eye out for any indications of penetration, like strange file access patterns or unapproved program installations. By examining data packets for questionable content or hostile behavior, network-based intrusion detection systems (IDS) thereafter monitor network traffic.

Intrusion detection systems (IDS) use a variety of methods to find and identify intrusions, such as anomaly-, signature-, and behavioral-based detection. Using pre-established criteria, signature-based detection detects possible threats by comparing known attack patterns or signatures against incoming traffic. On the other hand, anomaly-based detection creates a baseline of typical network activity and highlights any departures from it, possibly signifying an intrusion. The precision of its detection algorithms, the promptness of its notifications, and the capacity of security staff to react to possible threats are just a few of the variables that determine how successful IDS is.

1). Problem:

Traditional intrusion detection systems (IDS) frequently find it difficult to keep up with the sophisticated and dynamic nature of cyber threats in the quickly changing field of cybersecurity. Since rule-based signatures and pattern matching are the main components of the current IDS solutions, they are susceptible to new and undiscovered attack vectors. More intelligent and adaptable intrusion detection systems are desperately needed as the quantity and complexity of cyberthreats keep rising.

Machine Learning (ML) presents a promising avenue for enhancing the capabilities of intrusion detection systems. However, several challenges persist in the development and deployment of ML-based IDS. One of the primary issues is the scarcity of labeled datasets that accurately represent the diverse and constantly evolving cyber threat landscape. Moreover, the robustness and generalization capabilities of ML models in the face of adversarial attacks and false positives need to be addressed for effective real-world application.

2). Dataset:

The CIC-IDS2017 dataset is an important development in intrusion detection and prevention systems (IDSs) and IPSs. It addresses shortcomings in earlier datasets and satisfies eleven essential requirements for trustworthy benchmarking. The dataset contains realistic background traffic produced by the B-Profile system, a complete network configuration, and a variety of traffic with identified benign and attack flows. It shows the variety of attacks by capturing five days and including attacks such as Brute Force, DoS, DDoS, and more. The dataset records system calls and traffic from compromised machines, thereby guaranteeing heterogeneity. CIC-IDS2017 offers more than 80 network flow features that have been extracted, making it a strong platform for testing and developing anomaly-based intrusion detection methods in practical settings. We employed the CIC-IDS2017 dataset for our experiments.

3). Dataset Visualization and Normalization:

After loading the dataset we check the shape of the dataset. We also check the column name of the data to remove space from the column. We also check the datatype of each column in which we analyze, all column's data types are numeric except the label. Checking for missing values in a dataset is essential to ensure data quality, prevent biased analyses, and maintain model performance. We checked the missing values in the dataset and in each column, we found that there were total missing values are 5734 and dropped the all null occurrences. We also check the count of each label in which benign is 2271320 values. The dataset exhibits a varied distribution of cyber-attacks, ranging from highly prevalent to exceedingly rare occurrences. Attacks like DoS Hulk, PortScan, DDoS, and DoS GoldenEye dominate the dataset, reflecting their frequent appearance. Conversely, attacks such as Bot, Brute Force, XSS, Infiltration, Sql Injection, and Heartbleed are notably infrequent, suggesting their limited presence in the recorded incidents. This distribution underscores the differing frequencies of cyber threats within the dataset, offering insights into the prevalence and rarity of various attack types.

4). Binary Classification Conversation

In this project transforming a multi-class classification task into a binary one for effective model training and evaluation. To facilitate this, a new column named 'Attack' is introduced based on the existing 'Label' column in the dataset. The 'Attack' column serves to distinguish between normal network traffic, labeled as 'BENIGN,' and various types of cyberattacks, each labeled differently. By establishing this binary classification framework, the objective is to leverage the capabilities of binary classification algorithms. This transformation enables the development of models specifically designed to discern between benign and malicious network activities, providing a focused approach to enhance the detection and prevention of cyber threats.

5). Frequency Distribution

After the conversation, we checked the frequency distribution of the dataset. In this analysis, we notice that the benign percentage is 80.3% and other attacks are 19.7% as illustrated below in Figure 1.

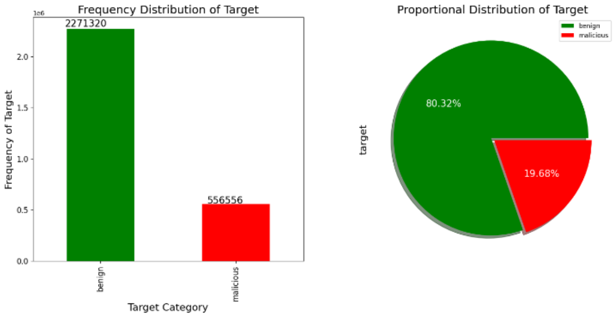


Figure 1: Frequency Distribution

If we observe the graph there is a vast difference in the frequencies of benign and malicious. Even in the case of the proportions benign is occupying a major share compared to malicious.

We also check the percentage of missing values, unique values, and percentage of one category values, and type against each column.

IV. MODELS BUILDING

We split the dataset into 3 forms.

- 60% data was used for Training.
- 20% for Validation.
- 20% for Testing.

We selected 1,696,725 samples with 78 features for training, 565,575 samples with the same 78 features for testing, and an additional 565,576 samples for validation. We also remove the features that have only one category value because that is not good for model prediction. We remove Fwd Avg Bytes/Bulk, Fwd Avg Bulk Rate, Fwd Avg Packets/Bulk, Bwd Avg Bytes/Bulk, Bwd Avg Packets/Bulk, Bwd Avg Bulk Rate, Bwd URG Flags and Bwd PSH Flags due to this only one unique value.

1). Feature Selection

Feature scaling is a crucial step in the final phase of data preprocessing for Machine Learning. It involves standardizing the independent variables of a dataset and placing them within a specific range. Essentially, feature scaling narrows the range of variables, enabling a fair and meaningful comparison between them. We perform a chi-squared (chi2) test for feature selection which helps in identifying the most significant features by assessing the independence between the feature and the target. This statistical method ranks features based on their relevance to the target variable, allowing us to select the most influential ones for model training. It's particularly useful in scenarios where we aim to understand which features contribute most to distinguishing between classes in a classification problem. Figure 2 illustrates the chi-squared test against each feature.

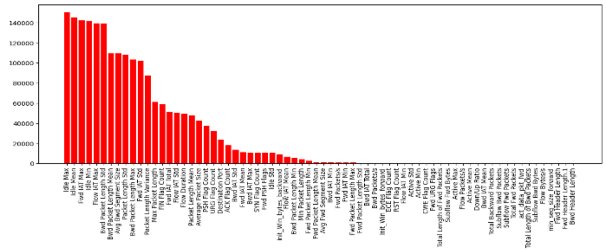


Figure 2: Chi-Squared

The importance of the features is represented in the graph and here we can observe that 28 features have very low importance. This visual representation helps us in feature selection effectively and precisely

We also check the cumulative feature score to select the number of features. As shown in Figure 3, the analysis graph shows that almost the first 40 features contained 99% of feature information. So, we select the value of K=40 which means we select the 40 feature.

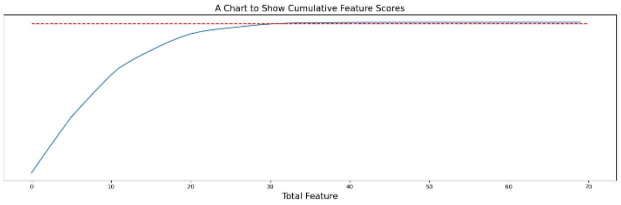


Figure 3: Cumulative Scores

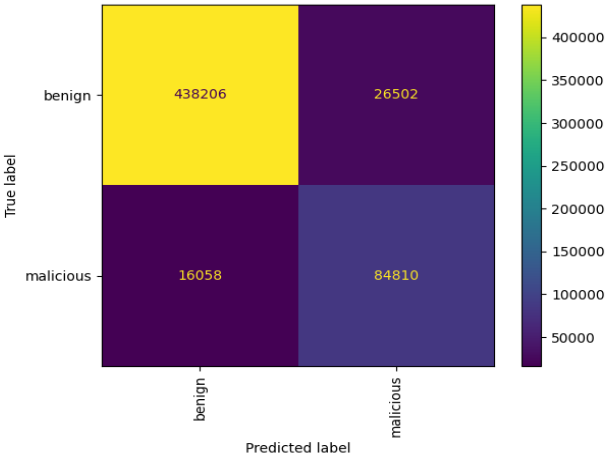
The figure represents the cumulative scores of the features. Only 40 features have 99% of the feature information so we can neglect the remaining features

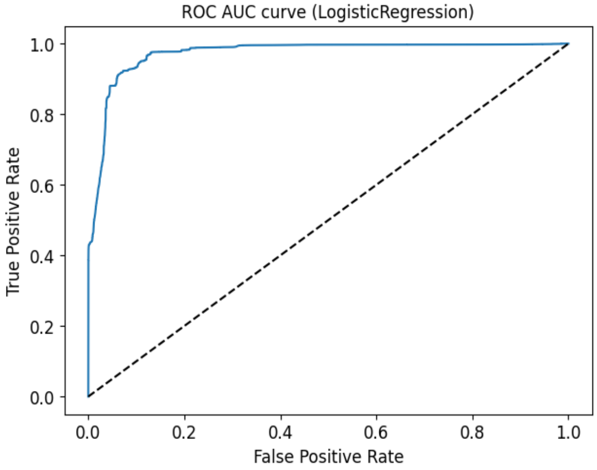
2). Training and Result

After selecting the features we use different single and hybrid models for binary classification of benign and malicious. The results of these classifiers are discussed in this section. The Performance metrics used to check their performance are accuracy, precision, recall, ROC-AUC (Receiver Operating Characteristic - Area under the Curve), and F1 score. We check accuracy on both training as well as testing. The F1 score shows the harmonic mean of precision and recall. Training Time is also considered, which is also an important parameter to check the performance of the model.

3). Logistic Regression

On Logistic Regression, the model performed healthy with a training accuracy of 92.45% and a testing accuracy of 92.47%. This suggests that the model generalizes effectively to new, unseen data. The model achieved a score of 76.19% in terms of precision. In recall, the model attained a score of 84.08%. The F1-score, an equilibrium between precision and recall, stands at 79.94%. The ROC-AUC score, a metric related to the model's capability to discriminate the classes, is 0.9717. A higher ROC-AUC score indicates better discrimination performance. A confusion matrix of Logistic regression and AUC is illustrated below.

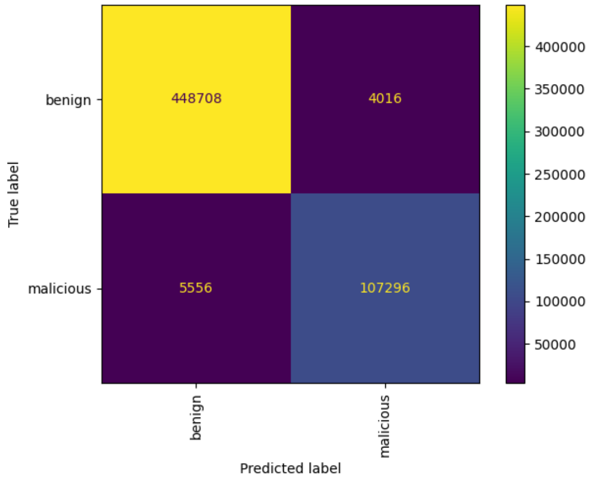


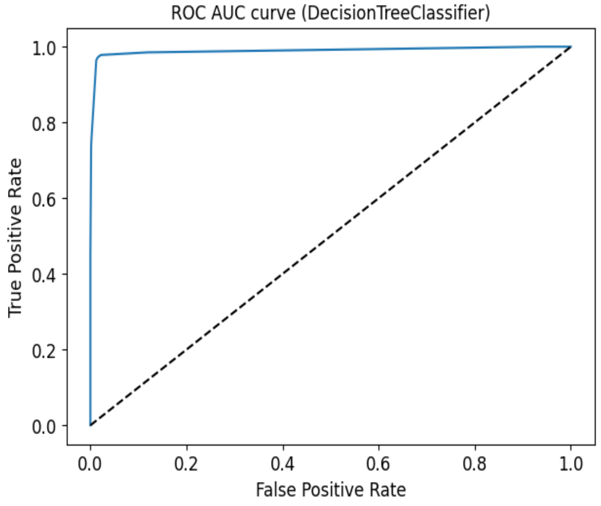


The above graphs represent the confusion matrix and ROC AUC curve of the Logistic regression model which helps in assessing the quality of the model. If we observe true values are larger than false values by a greater extent it indicates it has good accuracy.

4). Decision Tree

The model exhibits outstanding performance with a training accuracy of 98.29%. This capability extends to new, unseen instances, as evidenced by the testing accuracy of 98.31%. Precision, measuring the accuracy of positive predictions, stands at an impressive 96.39%, indicating a low rate of false positives. A recall of 95.08% showcases the model's effectiveness in capturing actual positive instances. The F1-Score, harmonizing precision and recall, is 95.73%, emphasizing a well-balanced performance. Moreover, the ROC-AUC score of 0.9897 underlines the model's excellent discriminatory ability, affirming its proficiency in distinguishing between positive and negative classes. The overall performance of the Decision Tree in the form of confusion matric and ROC curve is illustrated below.

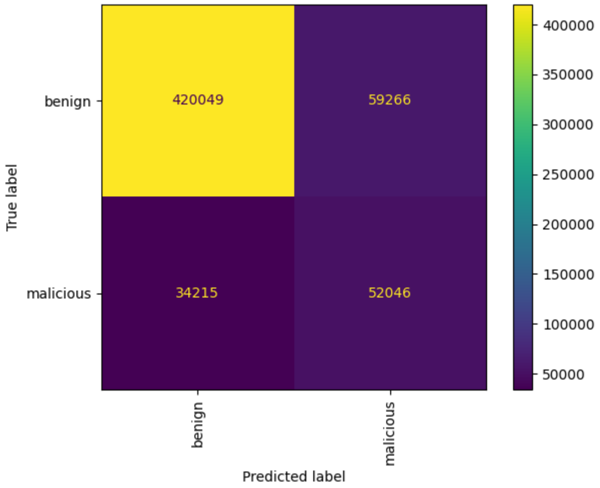


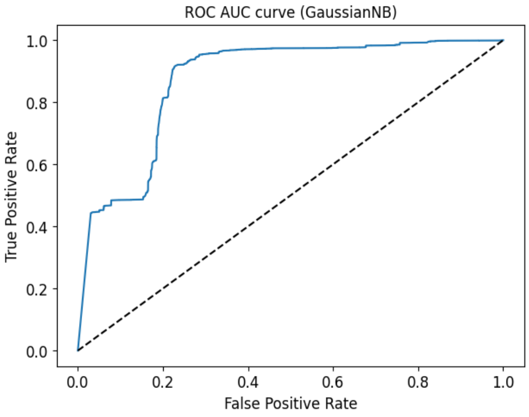


The above graphs represent the confusion matrix and ROC AUC curve of the Decision Tree classifier model which helps in assessing the quality of the model. If we observe true values are larger than false values by a greater extent it indicates it has good accuracy

5). Gaussian Naive Bayes

The Gaussian Naive performance demonstrated training accuracy of 83.49% indicating a reasonable ability to learn patterns from the training dataset. This competence extends to the testing dataset, with a corresponding accuracy of 83.47%, suggesting the model's capability to generalize to new, unseen data. Precision, measuring the accuracy of positive predictions, stands at 46.76%, indicating a relatively higher rate of false positives. The recall, representing the model's ability to capture actual positive instances, is at 60.34%, showcasing moderate proficiency. The F1-Score, balancing precision and recall, is 52.69%, suggesting a fair compromise between precision and recall. The ROC-AUC score of 0.8721 reflects a reasonable discriminatory ability, indicating a moderate capacity to distinguish between positive and negative classes.

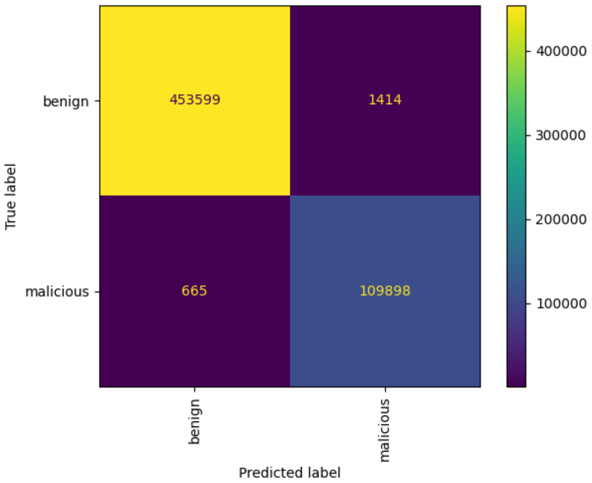


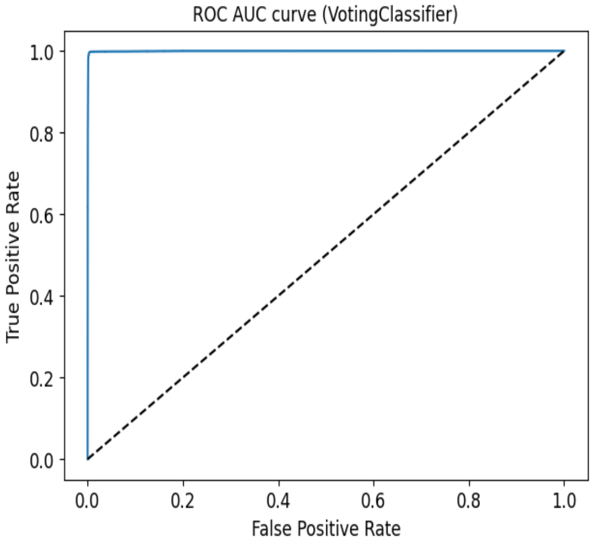


The above graphs represent the confusion matrix and ROC AUC curve of the Gaussian Naive Bayes model which helps in assessing the quality of the model. If we observe true values are larger than false values by some extent it indicates it has normal accuracy where we can improve it.

6). Hybrid Model (RF+LR)

A hybrid model is the combination of two models. In this, we combine random forest and logistic regression to check results. It takes more time than a single model but the model exhibits exceptional performance with a training accuracy of 99.65%. This competence extends seamlessly to the testing dataset, where the accuracy remains impressively high at 99.63%. The precision of 98.73% indicates an outstanding accuracy of positive predictions, minimizing false positives. The recall, measuring the model's proficiency in capturing actual positive instances, is remarkable at 99.4%, reflecting a high level of sensitivity. The F1-Score, harmonizing precision and recall, stands at an impressive 99.06%, highlighting the model's overall robustness. Additionally, the ROC-AUC score of 0.9995 signifies an almost perfect discriminatory ability, underlining the model's exceptional capacity to distinguish between positive and negative classes. These results collectively portray a highly effective and accurate model with minimal room for improvement. The following visualizations showcase the confusion matrix and ROC-AUC for the model's performance.

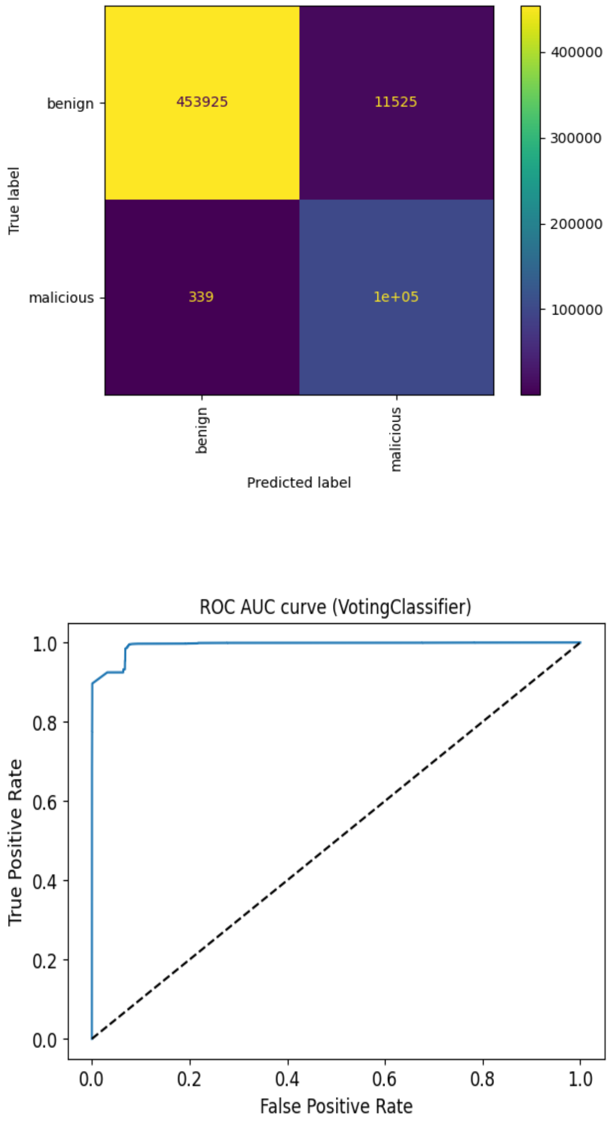




The above graphs represent the confusion matrix and ROC AUC curve of the Hybrid model combining Random forest and the logistic regression which helps in assessing the quality of the model. If we observe true values are larger than false values by a greater extent it indicates it has highest accuracy

7). Hybrid Model (DT-GNB)

In this hybrid model, we combine Decision Tree and Gaussian Naive Bayes. The model demonstrates strong performance metrics, with a training accuracy of 97.99%, indicating its proficiency in learning from the training dataset. This effectiveness carries over to the testing dataset, where the model maintains a high accuracy level of 97.9%. The precision of 89.65% reflects the model's ability to make accurate positive predictions while minimizing false positives. Notably, the recall is exceptionally high at 99.66%, highlighting the model's effectiveness in capturing the majority of actual positive instances. The F1-Score, a balanced measure of precision and recall, stands at 94.39%, indicating overall robust performance. The ROC-AUC score of 0.9934 further emphasizes the model's excellent discriminatory power, underlining its capability to distinguish between positive and negative classes. These results collectively underscore the model's effectiveness and reliability in classification tasks. The confusion matrix and ROC-AUC are illustrated below.



The above graphs represent the confusion matrix and ROC AUC curve of the Hybrid model combining the Decision Tree and Gaussian Naïve Bayes which helps in assessing the quality of the model. If we observe true values are larger than false values by a greater extent it indicates it has good accuracy

V. COMPARATIVE ANALYSIS

The exploration and development of a Hybrid Model represents a significant step in improving classification accuracy. This approach amalgamates various strengths from distinct classifiers, aiming to create a more robust and accurate prediction system. In our work, the creation of the Hybrid Model involves combining multiple classifiers, each contributing its unique approach to the final decision-making process. Leveraging the diverse algorithms of Logistic Regression, Decision Trees, and Gaussian Naive Bayes, the Hybrid Model harnesses the collective strengths of these models to enhance predictive power.

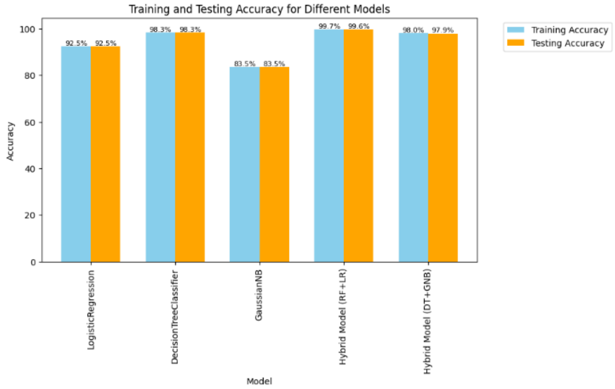
Model	**Training Accuracy**	**Testing Accuracy**	**Precision**	**Recall**	**F1**	**ROC-AUC**	**Training Time**
Logistic Regression	92.45	92.47	76.19	84.08	79.94	97.18	16.4778624
Decision Tree	98.29	98.32	96.39	95.08	95.73	98.97	16.436774
Gaussian Naïve Bayes	83.47	83.47	46.76	60.34	52.69	87.21	1.296666
Hybrid Model (RF+LR)	99.65	97.63	98.73	99.40	99.06	99.95	792.850931
Hybrid Model (DT+GNB)	97.99	97.90	89.65	99.66	94.39	99.34	65.857489

The above table compares the statistics of the all models we have applied till now. If we observe has better accuracies, precision, recall, ROC-AUC, and F1 values compared to others. But it takes time for training

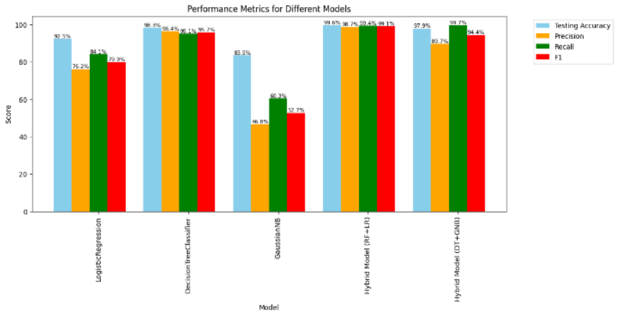
The novel aspect of our approach lies in the strategic combination of these classifiers, resulting in superior overall performance metrics. By blending the individual expertise of each model, the Hybrid Model achieves a commendable balance between precision, recall, accuracy, and area under the ROC curve, outperforming standalone models in many cases. This methodology showcases a synergistic utilization of various classifiers, demonstrating the potential for improved classification accuracy through a cohesive integration of diverse model capabilities. The table summarizes the performance of multiple machine learning models, such as Logistic Regression, Decision Tree Classifier, Gaussian Naive Bayes, and two hybrid models—each combining different algorithms. The metrics measured include Training and Testing Accuracy, Precision, Recall, F1 Score, ROC AUC Score, and Training Time.

Important Observation: -

Notably, the models demonstrated varied performance. Logistic Regression and Decision Tree Classifiers exhibited strong accuracy and balanced precision-recall scores. In contrast, Gaussian Naïve Bayes displayed lower accuracy but reasonable precision and recall. The hybrid models, which combined different classifiers, showcased superior performance across various metrics, indicating potential enhancements in predictive capabilities by combining algorithms. Among the evaluated models, the hybrid model formed by combining Random Forest with Logistic Regression demonstrated the highest performance across multiple metrics. This hybrid model exhibited exceptional accuracy, precision, recall, F1 Score, and ROC AUC Score, surpassing the individual models and the other hybrid combination. Hence, based on the measured metrics, the hybrid model of Random Forest and Logistic Regression appears to be the most effective in this context. Training testing accuracy of different models and performance metrics of all used model is illustrated below in figures.



The above graph compares the testing and training accuracies of all five models. If we observe the Hybrid model of Random Forest and Logistic regression it has highest accuracies and gaussian NB has the lowest accuracies



The above graph compares the testing accuracy, precision, Recall, and F1 values of the all five models.If we oberve hybrid model of Logistic regression and random forest has the highest values and gaussian NB has the lowest values.

VI. CONLUSION

This project focused on Intrusion Detection Systems (IDS) through Machine Learning (ML) approaches, addressing limitations in traditional IDS for dynamic cyber threats. Leveraging the CIC-IDS2017 dataset, robust preprocessing and model building were conducted, evaluating models based on key metrics. Logistic Regression and Decision Tree exhibited strong performance, emphasizing accuracy and balanced precision-recall trade-offs. Gaussian Naive Bayes, while showing lower accuracy, contributed reasonable precision and recall. Hybrid models, particularly the Random Forest and Logistic Regression combination (RF+LR), outperformed individual models, showcasing enhanced predictive capabilities. The comparative analysis highlighted the strategic integration of classifiers in hybrid models, emphasizing the potential for improved accuracy. The research contributes to fortifying cybersecurity systems, showcasing the effectiveness of ML-based IDS, particularly through hybrid models. Future work could explore additional datasets and emerging ML algorithms for further enhancements.

VII. REFERENCES

[1] Ravi, V., Chaganti, R., & Alazab, M. (2022). Recurrent deep learning-based feature fusion ensemble meta-classifier approach for intelligent network intrusion detection system. *Computers and Electrical Engineering*, *102*, 108156.

[2] Saba, T., Rehman, A., Sadad, T., Kolivand, H., & Bahaj, S. A. (2022). Anomaly-based intrusion detection system for IoT networks through deep learning model. Computers and Electrical Engineering, 99, 107810.

[3] Yadav, N., Pande, S., Khamparia, A., & Gupta, D. (2022). Intrusion detection system on IoT with 5G network using deep learning. Wireless Communications and Mobile Computing, 2022, 1-13.

[4] Logeswari, G., Bose, S., & Anitha, T. (2023). An intrusion detection system for sdn using machine learning. Intelligent Automation & Soft Computing, 35(1), 867-880.

[5] Musleh, D., Alotaibi, M., Alhaidari, F., Rahman, A., & Mohammad, R. M. (2023). Intrusion Detection System Using Feature Extraction with Machine Learning Algorithms in IoT. Journal of Sensor and Actuator Networks, 12(2), 29.

[6] Farooq, M. S., Abbas, S., Sultan, K., Atta-ur-Rahman, M. A., Khan, M. A., & Mosavi, A. (2023). A fused machine learning approach for intrusion detection system.

[7] Tcydenova, E., Kim, T. W., Lee, C., & Park, J. H. (2021). Detection of adversarial attacks in AI-based intrusion detection systems using explainable AI. Human-Centric Comput Inform Sci, 11.

[8] Hsu, C. Y., Wang, S., & Qiao, Y. (2021). Intrusion detection by machine learning for multimedia platform. Multimedia Tools and Applications, 80(19), 29643-29656.

[9] Martindale, N., Ismail, M., & Talbert, D. A. (2020). Ensemble-based online machine learning algorithms for network intrusion detection systems using streaming data. Information, 11(6), 315.