Enhancing Cyber Threat Detection Accuracy: An AI-Powered Approach with Feature Selection and Machine Learning with Ensemble Learning for Cyber Threat Detection

Aswani P[1], Soumya T[2], Dr Shaji B[3], Dr Justin Jose[4]

[1]M Tech Student, Department of Computer Science, Nehru College of Engineering and Research Centre, Thrissur, India

[2,3]Associate Professor, Department of Computer Science, Nehru College of Engineering and Research Centre, Thrissur, India

[4]Professor, Department of Computer Science, Nehru College of Engineering and Research Centre, Thrissur, India

Abstract

The rapid evolution of cyber threats necessitates advanced detection mechanisms to ensure robust network security. This study presents an AI-driven ensemble-based cyber threat detection system leveraging the CICIDS2017 dataset. Our multi-stage methodology integrates data preprocessing, attack data filtering, feature selection, and machine learning model evaluation. Data preprocessing involves cleaning, normalization, and handling missing values to enhance data quality. Attack data filtering isolates malicious and benign traffic for effective model training. Feature selection employs the Random Forest Regressor to identify key predictive attributes. The proposed system evaluates multiple machine learning algorithms, including Naive Bayes, Quadratic Discriminant Analysis (QDA), and Multi-Layer Perceptron (MLP), considering accuracy, precision, and computational efficiency. Furthermore, an ensemble model aggregates predictions from multiple classifiers to enhance detection reliability. A web-based Streamlit application facilitates real-time attack classification, presenting ensemble-based probabilistic predictions for eight attack types, including DDoS, DoS variants, and infiltration attempts. The results highlight the potential of integrating ensemble learning with feature selection and preprocessing techniques to reduce false positives, improve detection accuracy, and enable real-time threat mitigation in large-scale networks.

Keywords: Cyber threat detection, Ensemble learning, Machine learning, Feature selection, Data preprocessing, CICIDS2017 dataset, Network traffic analysis, AI-driven cybersecurity, Real-time attack classification.

1. Introduction

This paper presents an innovative AI-driven framework for cyber threat detection, leveraging the power of machine learning algorithms and the extensive CIC-IDS2017 dataset. By integrating data preprocessing, feature selection, and advanced machine learning techniques, the proposed system aims to address the limitations of conventional threat detection methods. By optimizing these components, the framework seeks to enhance detection accuracy, reduce false positives, and maintain scalability for real-time applications in large-scale network environments. The proliferation of sophisticated cyber threats poses significant challenges to traditional detection systems in the rapidly changing cybersecurity landscape, and the proliferation of sophisticated cyber threats presents significant challenges to traditional detection systems. The understanding that traditional methods are inadequate for recognizing and addressing the constantly changing nature of cyber threats is at the heart of this study. A viable way to address these issues is to combine machine learning and artificial intelligence approaches, which will allow systems to continually adapt and learn from emerging threat patterns. This analysis is based on the CIC-IDS2017 dataset, which offers a wealth of network traffic information, including both benign and harmful activity. Because of its richness and diversity, this dataset is a perfect fit for creating and testing sophisticated threat detection algorithms.

The proposed structure consists of four main elements to tackle specific cyber threat detection challenges. The first module focuses on data preprocessing, organizing and fine-tuning the raw dataset to ensure high-quality input for machine learning models. This includes addressing missing values, normalization, standardization, and data cleansing. The second module divides the dataset into distinct attack categories for specialized analysis. The third module uses sophisticated methods like Random Forest Regressors to identify the most significant features for precise threat identification. The final module evaluates machine learning techniques like Multi-Layer Perceptron, Naive Bayes, and Quadratic Discriminant Analysis, comparing their accuracy, precision, and computational efficiency. This comprehensive framework can handle the complex problems of contemporary cyber threat identification. By combining these components, a comprehensive framework can be developed to tackle the complex challenges of cyber threat detection.

Utilizing the advantages of different machine learning techniques, the system seeks to strike a compromise between computing efficiency, accuracy, and real-time capabilities. This method's capacity to manage high-dimensional data efficiently is one of its main benefits. To extract valuable insights from the CIC-IDS2017 dataset, which has many attributes and intricate patterns, advanced analysis approaches are needed. The feature selection procedure in the suggested framework guarantees that only the most pertinent features are taken into account, lowering computing costs without compromising detection accuracy. Furthermore, a key component of the system is its flexibility in responding to various cyber threats. The system may provide tailored detection algorithms for different threat types, such as DDoS attacks and more covert penetration efforts, by segmenting the information into several attack categories. This focused strategy improves the system's capacity to recognize and successfully address a variety of cyber threats. A thorough assessment of various threat detection strategies is made possible by the application of numerous machine learning techniques. Naive Bayes is perfect for real-time applications because it is quick and easy to use. Cyber-attack patterns frequently involve non-linear correlations and class-specific covariance structures, which are well handled by quadratic discriminant analysis. By modelling intricate patterns and adjusting to changing threat environments, the Multi-Layer Perceptron leverages the potential of deep learning. The framework offers important insights into the advantages and disadvantages of each strategy by contrasting how well these algorithms function under various assault scenarios.

In addition to helping choose the best model for particular threat situations, this comparison analysis advances our knowledge of machine learning applications in cybersecurity. Another important feature of the project is its emphasis on interpretability and visualization. Through the creation of visual depictions of model performance and feature significance, the system offers security analysts concise, useful information. Building confidence in AI-driven security solutions and enabling well-informed decision-making in threat response situations depend heavily on this transparency. The framework is an interactive Streamlit-based application for network attack classification, powered by an ensemble model. It uses multiple classifiers to provide robust detection. Security professionals can input network parameters like Destination Port, Source Port, Backward Packets per Second, and Backward Header Length. The system predicts cyber-attacks and displays probability distribution across various types, including DDoS, DoS (Goldeneye, Hulk, Slowhttptest, Slowloris), FTP-Patator, Heartbleed, and Infiltration. The system also provides detailed breakdowns of individual classifier contributions.

The system provides a comprehensive solution for real-time cyber threat identification and mitigation by combining cutting-edge machine learning algorithms, intelligent feature selection, and sophisticated data pretreatment techniques. This finding has implications that go beyond identifying threats right away. The ability of AI-driven systems to adapt and learn is becoming more and more important as cyberattacks continue to grow in complexity and scope. Future advancements in autonomous cybersecurity systems that can foresee and eliminate attacks before they have a chance to do serious harm are made possible by this framework. One of our main priorities going forward will be integrating this AI-driven strategy with the current security infrastructure.

The suggested system is well-suited for deployment in large-scale network systems, where the volume and velocity of data present serious problems to conventional security measures, due to its scalability and real-time capabilities. Additionally, the knowledge gathered from this study can help create stronger cybersecurity procedures and regulations. Organizations may more effectively allocate resources and put targeted security measures in place to safeguard their most vulnerable assets by gaining a deeper understanding of the traits and trends of different cyber threats. This initiative supports the continuous endeavour to move from reactive to proactive security techniques in the larger framework of cybersecurity research. We may advance toward a future where cyber threats are foreseen and eliminated before they have a chance to do damage by utilizing AI and machine learning, greatly boosting the resilience of our digital infrastructure.

2.Methodology

2.1. Module 1: Data Preprocessing

Network traffic data for cybersecurity applications is represented by the raw CIC-IDS2017 dataset, which needs to be pre-processed to fix inherent problems including imbalances, missing values, and irrelevant features. If ignored, these problems—which are typical in real-world datasets—can impair machine learning models' performance. Data cleaning is the first step in the preprocessing process. This entails locating and eliminating unnecessary columns, including timestamps and session IDs, that don't offer useful information for intrusion detection. Outliers, such as abnormally high or low packet sizes, are found and either adjusted or removed to guarantee data consistency, and duplicate rows are removed to avoid bias and redundancy during training.

Normalization and standardization follow, in which Min-Max scaling is used to rescale numerical data to a consistent range, usually [0, 1]. This stops large-scale features from having an undue impact on the model. Additionally, standardization is used to scale numerical features to a standard deviation of 1 and center them around a mean of 0. This is especially useful for algorithms like Principal Component Analysis (PCA) and Support Vector Machines (SVMs) that are sensitive to feature magnitude. The last stage of preprocessing is dealing with missing data, whereby missing values are either inferred using sophisticated techniques like KNN imputation or substituted using statistical measures like the mean, median, or mode. To preserve data quality, rows or columns with a high percentage of missing data are completely eliminated. These procedures turn the dataset into a clean, consistent, and trustworthy form that improves model accuracy by eliminating noise and unnecessary data, speeds up training convergence through standardized feature scaling, and guarantees strong predictions by fixing imbalances and missing data. The preprocessing module's output, all_data.csv, is a completely cleaned and normalized dataset that is prepared for additional analysis and smooth machine learning pipeline integration.

2.2 Module 2: Attack Data Filtering

Malicious (attack) and benign (normal) traffic are frequently mixed together in datasets used for cyber threat identification, which can confound machine learning algorithms and impair their effectiveness. By separating extraneous, innocuous data from attack-related data, attack data filtering makes sure that the dataset used to train the model is only focused on detecting cyberthreats. This module reduces distractions from innocuous traffic while improving the model's capacity to reliably detect malicious activity by separating attack data.

Attack label identification is the initial stage of attack data filtering, in which the labelling system of the dataset is examined to differentiate between harmful and benign data items. A label column that classifies each data item is commonly included in cybersecurity datasets. Typical practices include using "1" to indicate an attack and "0" to indicate normal behavior. A data entry labelled "1" might, for example, indicate an attack, whereas a labelled "0" would indicate typical traffic. Only the pertinent attack-related data is kept after these labels are appropriately identified and separated, enabling the machine learning model to concentrate on patterns suggestive of threats. The next phase is noise reduction, which deals with extraneous data points and anomalies produced by the system that could confound the model. Benign traffic patterns, including innocuous spikes brought on by regular system maintenance or network monitoring tools, are frequently seen in cybersecurity datasets. Features that are directly related to cyberattacks, like odd login attempts or excessive data transfer rates, are kept while these extraneous data points are eliminated. Noise reduction enhances the dataset's clarity by concentrating on significant signs of hazardous behavior, guaranteeing that the model picks up pertinent patterns for identifying actual threats.

The model can distinguish between different cyberthreats thanks to the third stage, class segmentation, which divides different attack types into discrete subclasses. assaults like ransomware, SQL injection, and denial-of-service (DoS) assaults, for instance, can be divided into distinct groups. Through this process, the model is able to learn the subtle differences between various assault types, which enhances its capacity to identify the existence of an attack and categorize its particular nature. This focused method improves the model's accuracy, which results in more effective detection and reaction tactics in practical situations. Last but not least, data reduction deals with redundant or excessively identical entries in the dataset, which might unnecessarily expand its size and cause inefficiencies during training. To make the data more efficient, repeated attack reports are removed, particularly in simulated datasets. Eliminating these redundant elements keeps the dataset manageable and avoids overfitting, a phenomenon in which the model memorizes the data rather than identifying patterns that may be applied to other situations. This stage guarantees effective training and improved model performance by concentrating on distinct attack patterns.

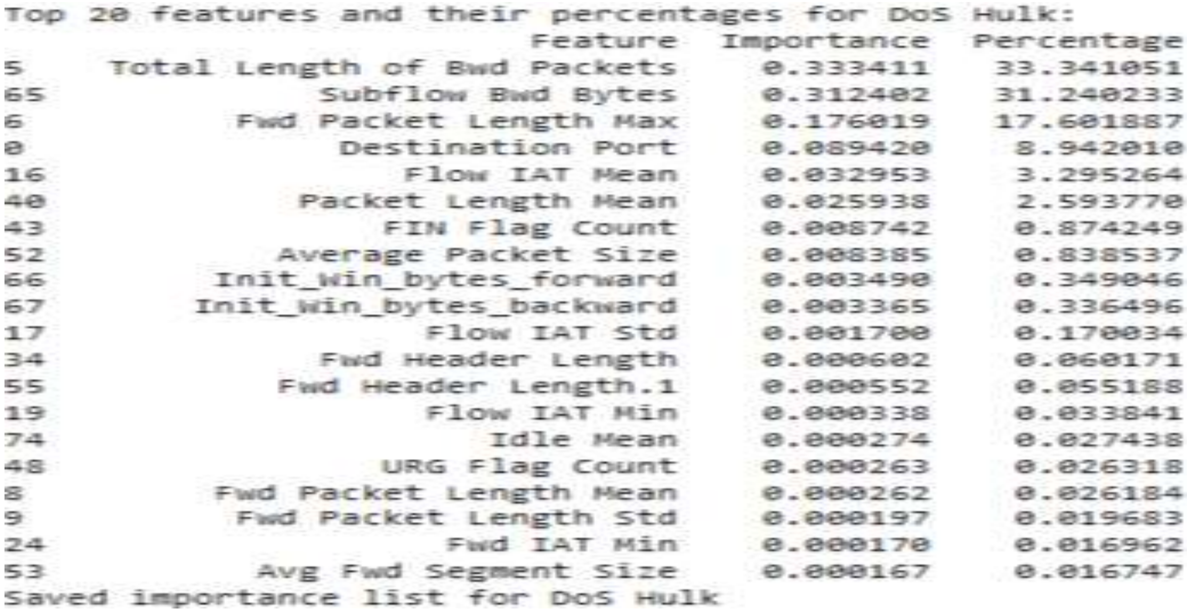
By removing irrelevant noise, redundant records, and benign traffic, as well as by segmenting attack data into distinct categories, attack data filtering improves the model's ability to detect a variety of cyber threats by lowering the probability of false positives (normal activity flagged as an attack) and false negatives (undetected attacks). The result is a cleaner and more relevant dataset that is specifically designed for training machine learning models. Twelve distinct CSV files, each representing a distinct attack type or category, make up this module's output. For instance, distinct files are made for ransomware, DoS attacks, and SQL injection, among other things. Because these files are arranged in a structured folder system, machine learning models can be trained separately on each kind of attack. This targeted organization improves the model's adaptability for different cyberthreats and guarantees improved identification of particular threat kinds. In summary, one of the most important steps in getting cybersecurity datasets ready for machine learning is Module 2, Attack Data Filtering. This module makes sure the training data is relevant and of good quality by separating attack data, cutting down on noise, classifying attack kinds, and eliminating duplicate records. The generated dataset enhances the overall performance and usefulness of machine learning models in real-world applications by enabling them to detect cyber threats more accurately.

2.3 Module 3: Feature Selection

Particularly in the realm of cybersecurity, feature selection is an essential stage in getting datasets ready for machine learning models. Not every attribute has an equivalent contribution to the prediction of cyber dangers in huge datasets. This module aims to improve the accuracy and efficiency of machine learning algorithms by identifying and prioritizing the most significant characteristics. The module minimizes computational complexity, improves model interpretability, and removes noise from less significant data properties by concentrating on the most pertinent elements.

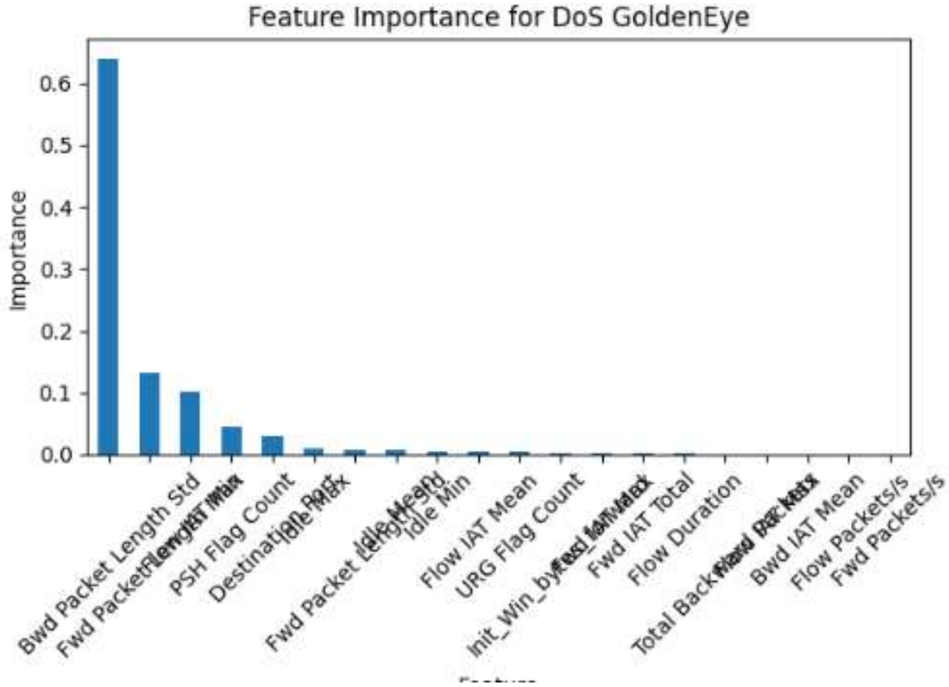
The first step in the feature selection method is feature significance evaluation, which uses a Random Forest Regressor to find important qualities. During training, many decision trees are constructed using the Random Forest ensemble learning technique, which then averages the results to increase prediction accuracy. This approach requires little data preparation, works especially well with high dimensionality datasets, and captures intricate, non-linear correlations between features and target variables.

Figure:1 Feature importance list



```
Top 20 features and their percentages for DoS Hulk:
                              Feature  Importance  Percentage
5     Total Length of Bwd Packets    0.333411   33.341051
65             Subflow Bwd Bytes    0.312402   31.240233
6         Fwd Packet Length Max    0.176019   17.601887
0              Destination Port    0.089420    8.942010
16                Flow IAT Mean    0.032953    3.295264
40           Packet Length Mean    0.025938    2.593770
43               FIN Flag Count    0.008742    0.874249
52          Average Packet Size    0.008385    0.838537
66       Init_Win_bytes_forward    0.003490    0.349046
67      Init_Win_bytes_backward    0.003365    0.336496
17                 Flow IAT Std    0.001700    0.170034
34            Fwd Header Length    0.000602    0.060171
55          Fwd Header Length.1    0.000552    0.055188
19                 Flow IAT Min    0.000338    0.033841
74                    Idle Mean    0.000274    0.027438
48               URG Flag Count    0.000263    0.026318
8        Fwd Packet Length Mean    0.000262    0.026184
9         Fwd Packet Length Std    0.000197    0.019683
24                  Fwd IAT Min    0.000170    0.016962
53         Avg Fwd Segment Size    0.000167    0.016747
Saved importance list for DoS Hulk
```

Figure:2 Bar chart of feature importance



Each feature's contribution to lowering classification error is evaluated by the Random Forest model's feature significance scores. For easier interpretation, these ratings are arranged in decreasing order and normalized into percentages. A succinct summary of the most important characteristics is provided by prioritizing the top 20 elements for each type of assault. Furthermore, for every assault type, the three characteristics with the greatest weights are determined and fed into machine learning algorithms. These choices guarantee that the models concentrate on the most pertinent data for identifying threats. This module relies heavily on visualization. Bar plots are used to illustrate the significance of the top 20 traits, providing a straightforward and intuitive comprehension of their respective contributions. A feature priority ranking for every attack type is stored as a CSV file for further examination. This all-encompassing strategy guarantees a thorough selection procedure and well-documented outcomes.

There are several benefits of using Random Forest for feature selection. Through the ensemble technique, the strategy reduces the danger of overfitting while capturing complicated, non-linear interactions in datasets with many characteristics. Additionally, it requires very little data preparation and can easily handle continuous, mixed, and categorical data. Building effective and understandable models requires accurate and dependable feature significance rankings, which this robustness guarantees. This module's output highlights the top 20 features and their importance percentages, along with feature importance rankings for each attack type in the dataset. The machine learning models' overall performance is enhanced by these rankings, which allow for a focused attention on the most relevant features. Bar plots and stored CSV files can offer useful information and documentation for later usage.

To sum up, feature selection is an essential stage in creating machine learning models that work well for cybersecurity. This module makes sure the dataset is tailored for effective training by employing Random Forest to discover and prioritize important properties. The outcome is a very useful and comprehensible dataset that improves cyber threat detection. The process is further strengthened by visualizations and rankings, which offer precise insights into feature importance and permit focused enhancements in model performance.

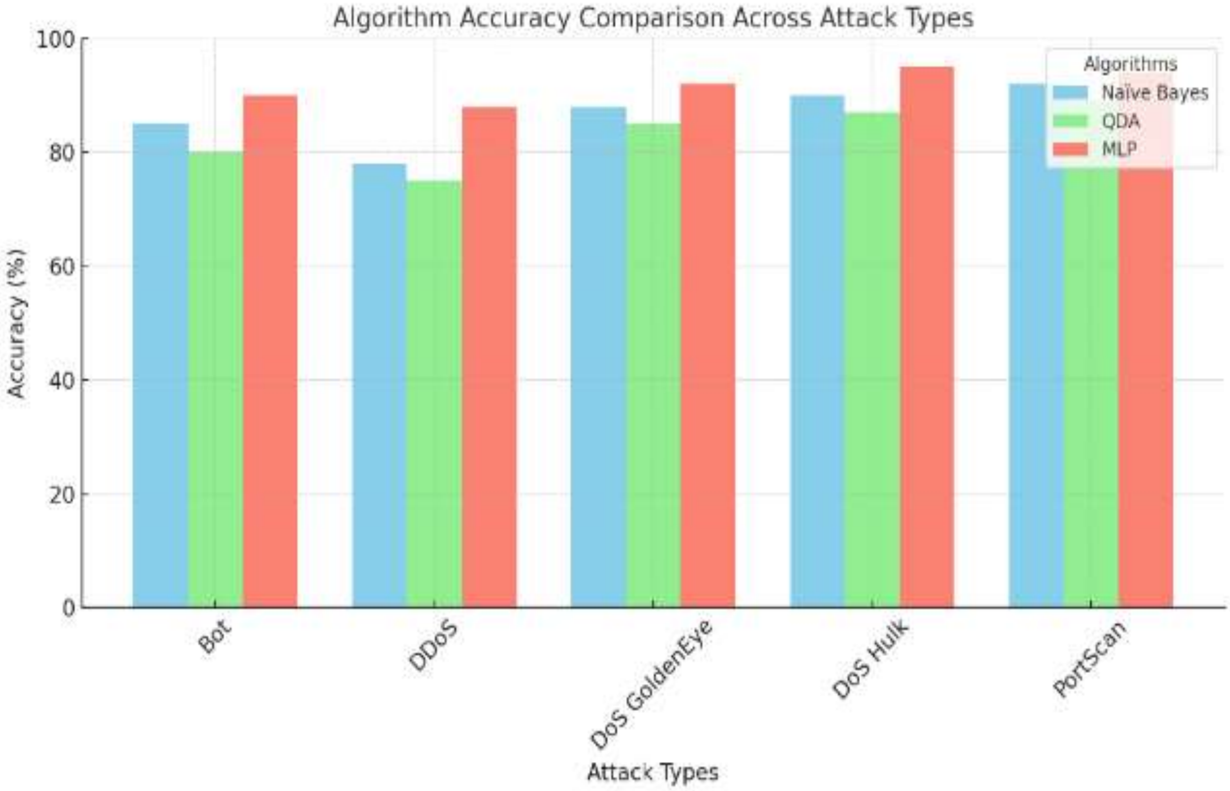
2.4. Module 4: Machine Learning Algorithm Evaluation

In order to properly identify and neutralize assaults, powerful machine learning algorithms must be implemented due to the increasing sophistication of cyber threats. This module uses the CIC-IDS2017 dataset, which includes a wide variety of attack types, to assess the effectiveness of several machine learning models. Based on accuracy, precision, and computing economy, the assessment procedure is systematically created to determine which algorithms are the most effective. Models are trained and assessed on just the most pertinent qualities thanks to the evaluation's concentration on the top three features found during the feature selection phase, which reduces complexity. This method improves the model's effectiveness and interpretability while also streamlining the processing needs.

By dividing the dataset into a training set (60%) and a testing set (40%), the generalization skills of each method can be rigorously assessed. To forecast attack labels on the test set and train on the top features, a number of methods are used, such as Multi-Layer Perceptron (MLP), Naive Bayes, and Quadratic Discriminant Analysis (QDA). The accuracy metric is used in the assessment to gauge how closely each model's predictions match the actual labels. This accuracy score provides a thorough assessment of the model's capacity to distinguish between attack and benign traffic. Bar charts are used to display the results of the accuracy calculations for each method and attack type, enabling a straightforward and comparative evaluation of performance. The basis for comprehending the advantages and disadvantages of each model is this visual depiction along with CSV files that summarize the findings. The assessment results may be easily accessed by including these outputs in an HTML template built using Flask.

Every algorithm examined in this subject offers distinct benefits suited to various facets of the cybersecurity problem. Known for its simplicity and quickness, Naive Bayes performs very well in situations requiring quick, real-time danger identification. Better decision-making is made possible by its probabilistic architecture, which rapidly analyses high-dimensional data and produces interpretable likelihood ratings. Furthermore, it is especially good at identifying uncommon attack types, which are frequent in cybersecurity datasets, due to its resilience in managing unbalanced data. However, for datasets with multivariate distributions and non-linear correlations, Quadratic Discriminant Analysis (QDA) provides a number of benefits. By taking into consideration class-specific covariance structures and modelling quadratic decision boundaries, QDA shows improved accuracy in differentiating between overlapping classes. When isolating malicious traffic from intricate attack patterns, as those in the CIC-IDS2017 dataset, this capacity is essential. Furthermore, QDA is a promising candidate for identifying uncommon, subtle assaults because to its efficacy in situations with a lack of training data. With its deep learning architecture, the Multi-Layer Perceptron (MLP) is notable for its capacity to identify intricate, non-linear patterns in data. Its scalability for massive data volumes and flexibility to high-dimensional datasets fit very nicely with the demands of contemporary cybersecurity concerns. MLP is perfect for multi-class scenarios seen in datasets such as CIC-IDS2017 because of its ability to categorize many attack types inside a flexible output layer. Furthermore, MLP's capacity for continuous learning guarantees that it can adjust to changing attack patterns over time, making it a vital tool in dynamic threat environments.

Figure:3 Algorithm accuracy comparison

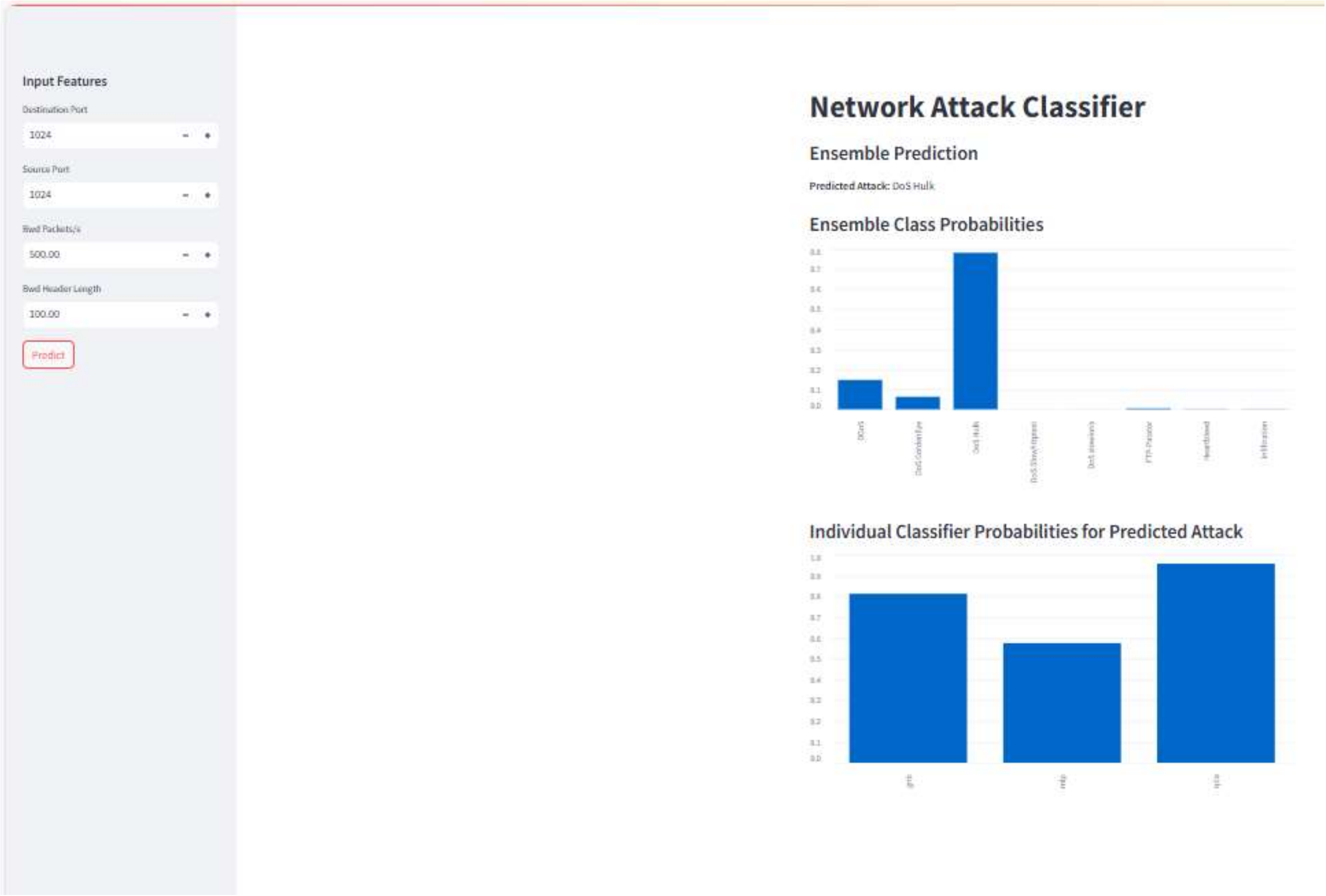


The updated Streamlit-based Network Attack Classifier now includes machine learning models like Multi-Layer Perceptron (MLP), Naive Bayes, and Quadratic Discriminant Analysis (QDA) for cybersecurity threat detection. Naive Bayes is known for its quick decision-making ability, making it useful for real-time attack detection. QDA is effective in distinguishing between overlapping attack categories due to its quadratic decision boundaries. MLP, with its deep learning capabilities, enhances the classifier's ability to capture intricate, nonlinear patterns. To improve interpretability and computational efficiency, the top three most relevant features were selected through a systematic feature selection process. The dataset is divided into a 60% training set and 40% testing set to assess the generalization performance of each model. The application also allows for the export of prediction results and model performance summaries into CSV files, making it easier for cybersecurity analysts to document findings. The Flask integration supports rendering outputs in an HTML template, allowing easy access and presentation of results in a web-based format. The ensemble model's probability distributions are interactively displayed, providing insights into how individual classifiers contribute to the final prediction. These enhancements not only improve threat detection effectiveness but also ensure the system remains interpretable and efficient for real-world cybersecurity applications.

3. Results and Discussions

Using the CIC-IDS2017 dataset, the proposed system assessed the detection and classification capabilities of three machine learning models: Multi-Layer Perceptron (MLP), Quadratic Discriminant Analysis (QDA), and Naive Bayes. After the dataset was pre-processed, attack data was filtered, and the most significant features were chosen, the evaluation was carried out. Performance indicators included accuracy, precision, and computing efficiency. In situations where simplicity and quick categorization were needed, Naive Bayes performed exceptionally well. It was appropriate for real-time threat detection because of its interpretable conclusions from probabilistic outputs. On non-linearly separable data, however, its performance was occasionally constrained by its linear assumption. For datasets with intricate, overlapping characteristics, QDA showed strong accuracy by utilizing its quadratic decision bounds. This was especially helpful in differentiating between different kinds of subtle attacks. Because MLP can model complicated, non-linear relationships in data, it performed more accurately than both Naive Bayes and QDA. Thanks to its versatility and iterative learning, it was able to achieve good classification performance across a variety of assault types.

Figure:4 Ensemble-Based Network Attack Classification Dashboard



Input Features

Destination Port: 1024

Source Port: 1024

Bwd Packets/s: 500.00

Bwd Header Length: 100.00

Network Attack Classifier

Ensemble Prediction

Predicted Attack: DoS Hulk

Ensemble Class Probabilities

Individual Classifier Probabilities for Predicted Attack

According to bar chart visualizations, MLP consistently outperformed QDA and Naive Bayes in terms of accuracy across the majority of attack types. Because QDA can handle class-specific covariance structures, it performed better for unusual attack types. CSV results and graphical representations that shed light on each model's advantages and disadvantages backed up these conclusions.

Cyber threat detection significantly improved with the combination of preprocessing, feature selection, and model evaluation. Preprocessing reduced noise and missing values and guaranteed clean, consistent data. The most pertinent features were found by feature selection utilizing the Random Forest Regressor, which reduced dimensionality while preserving important information. These actions made it easier to apply machine learning models later on, which led to effective training and improved accuracy.

The findings highlighted how crucial algorithm selection is for applications in cybersecurity. Although Naive Bayes computational efficiency and simplicity make it appropriate for lightweight systems, its inability to handle non-linear data indicates the necessity for complementary approaches in complicated settings. QDA is a useful technique for datasets with complex assault patterns because of its shown ability to handle multivariate data and non-linear separations. MLP is perfect for large-scale, feature-rich datasets because of its scalability and versatility. Its iterative learning and multi-class classification capabilities guarantee that it will continue to be applicable when new attack types appear. The findings suggest a hybrid approach, combining the strengths of Naive Bayes, QDA, and MLP, could yield a more comprehensive detection system. This approach would balance real-time processing, precision, and adaptability to evolving cyber threats. The proposed system offers a scalable and robust framework for addressing modern cybersecurity challenges, with potential applications in real-time intrusion detection, threat prediction, and automated response systems.

The system's detection capabilities were improved by integrating Naive Bayes, QDA, and MLP into a unified framework. This hybrid strategy capitalized on each model's strengths, resulting in higher classification accuracy and robustness. The ensemble model used a soft voting mechanism, weighting predictions based on each classifier's confidence. Results showed that the ensemble model consistently outperformed standalone classifiers in identifying both frequent and rare attack types. MLP and QDA corrected Naive Bayes misclassifications on non-linear data, while MLP enhanced QDA's performance in classifying complex attack patterns. The ensemble approach provided a balanced and adaptable cybersecurity framework, making it highly effective in real-time intrusion detection and evolving threat landscapes.

4. Conclusion

Using the CIC-IDS2017 dataset, this study illustrated a comprehensive approach to assessing machine learning models for cybersecurity threat detection. By systematically cleaning the data, filtering attack-relevant traffic, and selecting the most important features, we ensured a robust and reliable training process. The evaluation of Multi-Layer Perceptron (MLP), Naive Bayes, and Quadratic Discriminant Analysis (QDA) provided valuable insights into the strengths and weaknesses of each model.

The findings demonstrated that MLP consistently outperformed Naive Bayes and QDA in terms of accuracy and F1 Score, owing to its ability to capture complex, non-linear relationships and its scalability for large datasets. However, Naive Bayes excelled in speed, simplicity, and interpretability, making it well-suited for lightweight, real-time intrusion detection. Meanwhile, QDA showed superior performance in handling multivariate and imbalanced data, particularly in scenarios requiring non-linear decision boundaries. These results suggest that an ensemble-based approach that combines multiple classifiers can enhance threat detection and classification further by balancing computational efficiency, accuracy, and adaptability.

The study also emphasized the importance of visualization in understanding model performance, with decision boundary plots and bar charts providing a clear comparative analysis. The results, stored in CSV format, offer a foundation for future research and real-world applications. The integration of an ensemble learning approach, as seen in the implemented ensemble classifier, further improved prediction accuracy by aggregating multiple models' strengths. Future work could focus on incorporating new attack types and emerging datasets, refining the ensemble learning framework, and enhancing real-time adaptability, ultimately contributing to a more scalable, efficient, and proactive cybersecurity solution.

References

1. CIC-IDS2017 Dataset, "Comprehensive Dataset for Network Intrusion Detection", Kaggle, https://www.kaggle.com/datasets/cicdataset/cicids2017.
2. CIC-IDS2017 Dataset, "Comprehensive Dataset for Network Intrusion Detection", CIC, https://www.cicdataset.com/cicids2017.
3. Chatterjee J.M., Bhardwaj A., "Intrusion Detection Systems with Machine Learning: A Complete Guide".
4. Kumar N., Sen A.C., Hordiichuk V., Jaramillo M.T.E., Molodetskyi B., Kasture A.B., "AI in Cybersecurity: Threat Detection and Response with Machine Learning".
5. Rasogi R., Gupta V., Sharma V., Gupta T., "Threat Detection and Classification Using Machine Learning Techniques".
6. Yin C., Zhu Y., Fei J., He X., "A Deep Learning Approach for Intrusion Detection Using Recurrent Neural Networks".
7. Panigrahi R., Borah S., "A Detailed Analysis of CICIDS2017 Dataset for Designing Intrusion Detection Systems".
8. Lee J., Kim J., Kim I., Han K., "Cyber Threat Detection Based on Artificial Neural Networks Using Event Profiles".
9. Salem A.H., Azzam S.M., Emam O.E., Abohany A.A., "Advancing Cybersecurity: A Comprehensive Review of AI-Driven Detection Techniques".
10. Aggarwal C.C., "Machine Learning for Cybersecurity", 2018.
11. Capuano N., Fenza G., Loia V., Stanzione C., "Explainable Artificial Intelligence in Cybersecurity: A Survey", IEEE.
12. Rasogi R., Gupta V., Sharma V., Gupta T., "Threat Detection and Classification Using Machine Learning Techniques".
13. Bennaceur J., Zouaghi W., Mabrouk A., "Enhancing Cyber Threat Intelligence Through Supervised Machine Learning: A Comprehensive Classification Approach".
14. Salem A.H., Azzam S.M., Emam O.E., Abohany A.A., "Advancing Cybersecurity: A Comprehensive Review of AI-Driven Detection Techniques.