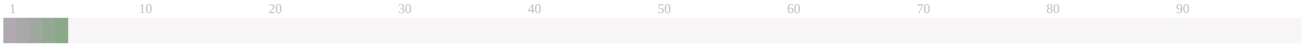
The Report is Generated by DrillBit Plagiarism Detection Software

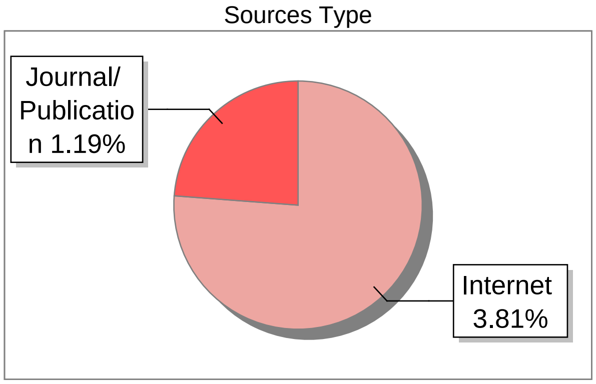
Submission Information

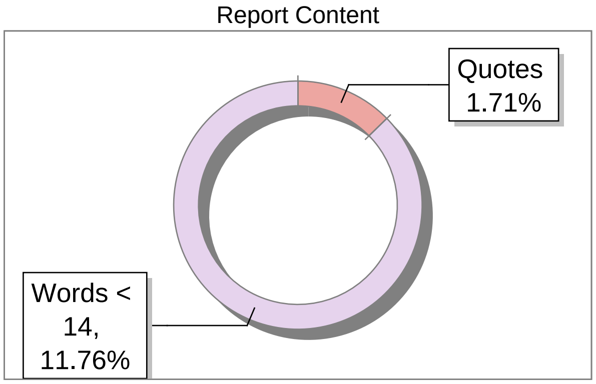
Author Name	Misba Arshad
Title	Cybersecurity Threats:Machine Learning Strategies for Cyber Attack Detection
Paper/Submission ID	3577181
Submitted by	premu.kumarv@gmail.com
Submission Date	2025-05-05 14:13:39
Total Pages, Total Words	13, 2518
Document type	Research Paper

Result Information

Similarity **5 %**







Exclude Information

Quotes	Excluded
References/Bibliography	Excluded
Source: Excluded < 14 Words	Excluded
Excluded Source	**0 %**
Excluded Phrases	Not Excluded

Database Selection

Language	English
Student Papers	Yes
Journals & publishers	Yes
Internet or Web	Yes
Institution Repository	Yes

A Unique QR Code use to View/Download/Share Pdf File

DrillBit

DrillBit Similarity Report

5	5	A
SIMILARITY %	MATCHED SOURCES	GRADE

A-Satisfactory (0-10%)
B-Upgrade (11-40%)
C-Poor (41-60%)
D-Unacceptable (61-100%)

LOCATION	MATCHED DOMAIN	%	SOURCE TYPE
1	dl.acm.org	1	Internet Data
2	www.leewayhertz.com	1	Internet Data
3	arxiv.org	1	Publication
4	translate.google.com	1	Internet Data
5	www.techscience.com	1	Internet Data

Cybersecurity Threats: Machine Learning Strategies for Cyber Attack Detection

Misba Arshad
Maseeka Koinain
BE Student
TOCE, Bangalore

Abstract

Data integrity, financial stability, and operational continuity are seriously threatened by the growing frequency of cyberattacks in networked systems. In order to classify network traffic as either benign or malicious, this study uses the CICIDS2017 dataset to build machine learning models such as Random Forest, XGBoost, and Linear Support Vector Classifier (Lin-earSVC).
Precision, recall, and F1-score are used to assess the models, and bar plots for training and testing datasets, a scatter plot for testing metrics, a feature correlation heatmap, and confusion matrices for each model are used to visualize performance. In order to support strong cybersecurity frameworks and intelligent intrusion detection systems, the study intends to improve attack detection accuracy and offer practical insights into network behavior.

Keywords: intrusion detection, binary classification, precision, recall, F1-score, cybersecurity, real-time detection, feature correlation, confusion matrix, machine learning,
random forest, XGBoost, LinearSVC, CICIDS2017 dataset, and cyber attack detection.

Introduction

I.1 Background and Motivation

The risk of cyberattacks, such as Distributed Denial of Service (DDoS), brute force attacks, and botnets, which jeopardize data security and interfere with operations, has increased due to the quick growth of digital networks. By 2025, cybercrime expenses are expected to surpass trillions of dollars annually, according to global cybersecurity reports, underscoring the urgent need for sophisticated intrusion detection systems. Protecting digital infrastructure, guaranteeing service dependability, and building trust in networked systems—especially in urban and business contexts—all depend on the efficient detection of malicious network traffic.

I.2 The Role of Machine Learning in Cyber Attack Detection

Conventional intrusion detection systems (IDS) frequently use rule-based or signature-based strategies, which are unable to keep up with new and advanced cyberthreats. By identifying intricate patterns in vast datasets, machine learning (ML) offers a potent substitute that makes predictive and adaptive detection possible. ML models are perfect for real-time cybersecurity applications because they can recognize irregularities and categorize network traffic according to a variety of characteristics. This study improves network security in dynamic environments by using machine learning (ML) to move from reactive to proactive intrusion detection.

I.3 Dataset Used

II The CICIDS2017 dataset, a network intrusion detection benchmark created by the Canadian Institute for Cybersecurity, is used in this study. The dataset records actual network traffic over several days, including both benign activity and different types of attacks (like DDoS and SQL Injection). It is the perfect tool for training and assessing machine learning models for cyberattack detection because it offers comprehensive features like packet counts, flow durations, and protocol types.

II.1 Machine Learning Models for Cyber Attack Detection

We investigate three machine learning models for binary classification of network traffic:

- **Random Forest**: An ensemble method that combines multiple decision trees to improve classification accuracy.

- **XGBoost**: A gradient boosting framework known for its efficiency and performance in handling imbalanced data.

- **Linear Support Vector Classifier (LinearSVC)**: A linear classifier optimized for high-dimensional data, capable of distinguishing between benign and malicious traffic.

Each model is evaluated for its ability to detect attacks, classify traffic, and support real-time intrusion detection systems.

II.2 Evaluation Metrics

The models are assessed using the following metrics:

- **Precision**: Measures the proportion of correctly identified malicious traffic among all predicted malicious instances.
- **Recall**: Assesses the ability to detect all actual malicious traffic instances.
- **F1-Score**: The harmonic mean of precision and recall, providing a balanced mea- sure of model performance.

These metrics ensure a thorough evaluation of detection accuracy and robustness.

II.3 Research Objectives and Contributions

This study is guided by three primary objectives:

1. To develop machine learning models for accurate detection of cyber attacks in network traffic.
2. **To conduct a comparative analysis of Random Forest, XGBoost, and LinearSVC for intrusion detection.**
3. To provide data-driven insights for enhancing network security and supporting in- telligent intrusion detection systems.

The contributions include a scalable ML framework for cyber attack detection, compre- hensive performance visualizations, and actionable insights for network security planning.

III Literature Survey

The development of intrusion detection systems has been greatly impacted by recent developments in machine learning. In order to overcome the difficulties of real-time malicious network traffic detection, numerous studies have investigated a variety of models and datasets.

Using the CICIDS2017 dataset, Alsaedi et al. (2020) demonstrated a thorough method for intrusion detection. They highlighted the significance of feature selection to lower computational overhead and showed how ensemble techniques like Random Forest can improve detection accuracy. Their research made clear that cybersecurity applications require scalable machine learning models.

In order to detect anomalies in network traffic, Kaur et al. (2019) presented a deep learning framework that makes use of Recurrent Neural Networks (RNNs). Their model achieved greater precision and recall in attack classification tasks by utilizing temporal dependencies, surpassing conventional techniques such as Support Vector Machines (SVM) and Decision Trees.

The use of XGBoost for intrusion detection was investigated by Ferrag et al. (2020), who noted that it could manage unbalanced datasets such as CICIDS2017. Their tests demonstrated that XGBoost offers excellent efficiency and accuracy, which makes it appropriate for real-time detection in extensive networks.

Ahmad et al. (2022) also looked into the application of Support Vector Classifiers to the detection of cyberattacks. Their research demonstrated how well linear SVC works in high-dimensional spaces, especially when paired with subsampling strategies to shorten training times without sacrificing detection accuracy.

According to this survey, sophisticated approaches that make use of ensemble methods, gradient boosting, and optimized classifiers are more successful for intricate intrusion detection tasks, even though traditional ML models provide baseline performance. Our suggested system is based on these studies and contrasts Random Forest, XGBoost, and LinearSVC using the CICIDS2017 dataset.

IV Equations

This study employs several mathematical equations to evaluate the performance of our cyber attack detection models. These metrics quantify the accuracy and reliability of the models classifications.

IV.1 Precision

Precision measures the accuracy of positive predictions, i.e., the proportion of correctly identified malicious traffic among all predicted malicious instances. It is defined as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

where:

- TP = True Positives (correctly predicted malicious traffic),
- FP = False Positives (benign traffic incorrectly predicted as malicious).

IV.2 Recall

Recall measures the ability of the model to detect all actual malicious traffic. It reflects how well the model captures attacks when they occur:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

where:

- FN = False Negatives (malicious traffic that the model failed to predict).

IV.3 F1-Score

F1-Score is the harmonic mean of precision and recall, providing a balanced measure of model performance:

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

This metric is particularly useful for imbalanced datasets like CICIDS2017, where malicious traffic is less frequent than benign traffic.

These metrics collectively provide a comprehensive framework to assess and compare the performance of the models developed in this project.

V Limitations in Existing Systems

Traditional intrusion detection systems (IDS) face several limitations that hinder their effectiveness in todays dynamic and threat-rich network environments.

V.1 Lack of Real-Time Adaptability

Most conventional IDS rely on predefined signatures or static rules, which cannot adapt to new or evolving attack patterns in real time. This rigidity leads to delayed detection and increased vulnerability to zero-day attacks.

V.2 Inability to Handle Complex Patterns

Network traffic exhibits non-linear and high-dimensional patterns, which traditional al- gorithms like rule-based systems or basic statistical methods struggle to capture. As a result, their detection accuracy is often compromised, especially for sophisticated multi- stage attacks.

V.3 Limited Use of Available Data

Many existing systems rely on a narrow set of features, such as packet headers or traffic volume, without leveraging the rich, multidimensional data available from modern net- work monitoring tools. This underutilization limits their ability to detect subtle attack signatures.

V.4 Scalability Issues

As networks grow in size and complexity, traditional systems face scalability challenges. Many are not designed to process large-scale data efficiently, leading to performance degradation and increased latency in attack detection.

V.5 Lack of Predictive Intelligence

Conventional IDS are predominantly reactive, focusing on identifying attacks after they occur rather than predicting and preventing them. This reactive approach results in prolonged exposure to threats and increased damage.

These limitations highlight the need for advanced, intelligent systems that can learn from data, adapt in real time, and provide predictive insights for cyber attack detection.

VI Proposed System Statement

We suggest a machine learning-based framework for identifying cyberattacks using historical and current network traffic data in order to overcome the shortcomings of current intrusion detection systems. Our system's ability to distinguish between malicious and benign traffic allows for proactive cybersecurity measures.

To determine the best strategy for attack detection, the main goal is to create and evaluate three predictive models: Random Forest, XGBoost, and Linear Support Vector Classifier (LinearSVC).The system analyzes characteristics like flow duration, packet counts, and protocol types using the CICIDS2017 dataset in order to identify malicious patterns. LinearSVC offers a linear classification method optimized for high-dimensional spaces, Random Forest uses ensemble learning to increase classification accuracy, and XG-Boost handles imbalanced data with gradient boosting. Performance is maximized by sampling the dataset (20% of records), subsampling it further (10%) for LinearSVC training, and reducing dimensionality to 20 key features through feature selection.

Precision, recall, and F1-score are used by the system to assess performance. The results are displayed using bar plots for training and testing datasets, a scatter plot for testing metrics, a feature correlation heatmap, and confusion matrices for every model. The ultimate goals of the suggested system are to improve network security, lessen the vulnerability of attacks, and facilitate intelligent intrusion detection.

VII Methods and Algorithms Used

This section outlines the machine learning models and algorithms used in our proposed system, selected for their ability to detect malicious patterns in network traffic.

VII.1 Random Forest

Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the majority vote for classification. It is robust to overfitting and effective for high-dimensional data, making it suitable for detecting complex attack patterns in network traffic.

VII.2 XGBoost

XGBoost (Extreme Gradient Boosting) is a scalable gradient boosting framework that optimizes performance through parallelization and regularization. It excels in handling imbalanced datasets like CICIDS2017, where benign traffic dominates, and provides high accuracy in attack classification.

VII.3 Linear Support Vector Classifier (LinearSVC)

LinearSVC is a linear classifier optimized for high-dimensional data. It uses a linear kernel to separate benign and malicious traffic, trained on a subsampled dataset (10% of training data) to reduce computational complexity while maintaining effectiveness in attack detection.

VII.4 Data Preprocessing

The following preprocessing steps were applied:

- **Normalization**: Features were scaled using StandardScaler to ensure consistent ranges.
- **Missing Value Handling**: NaN and infinite values were replaced with zeros.
- **Feature Selection**: Reduced to 20 features using Random Forest importance scores.
- **Data Splitting**: 80% training, 20% testing, with stratification to preserve class distribution.

VII.5 Model Training

Each model was trained on the preprocessed CICIDS2017 dataset. Hyperparameters (e.g., Random Forests $n_{e}stimators = 50, XGBoostsmax_{d}epth = 6, LinearSVCsC = 0.1)$*wereselectedtobalanceaccuracyandcomputationalefficiency.Performancewasevaluatedusingpr score, withvisualizationsprovidinginsightsintomodelbehavior.*

VIII Advantages of the Proposed System

The proposed machine learningbased cyber attack detection system offers several advan- tages over traditional intrusion detection methods.

VIII.1 Improved Detection Accuracy

By leveraging advanced ML models like Random Forest, XGBoost, and LinearSVC, the system achieves higher detection accuracy. These models capture complex patterns in network traffic, improving the identification of malicious activities.

VIII.2 Real-Time Detection Capability

The system processes large volumes of traffic data efficiently, enabling real-time detection. This allows network administrators to respond promptly to threats, minimizing damage and downtime.

VIII.3 Adaptability to Evolving Threats

Unlike signature-based systems, the proposed ML models adapt to new attack patterns by learning from data. This adaptability is crucial in dynamic network environments where threats evolve rapidly.

VIII.4 Comparative Model Evaluation

By comparing multiple algorithms, the system provides insights into their strengths and limitations, allowing for the selection of the most suitable model for specific network scenarios.

VIII.5 Scalability for Network Security

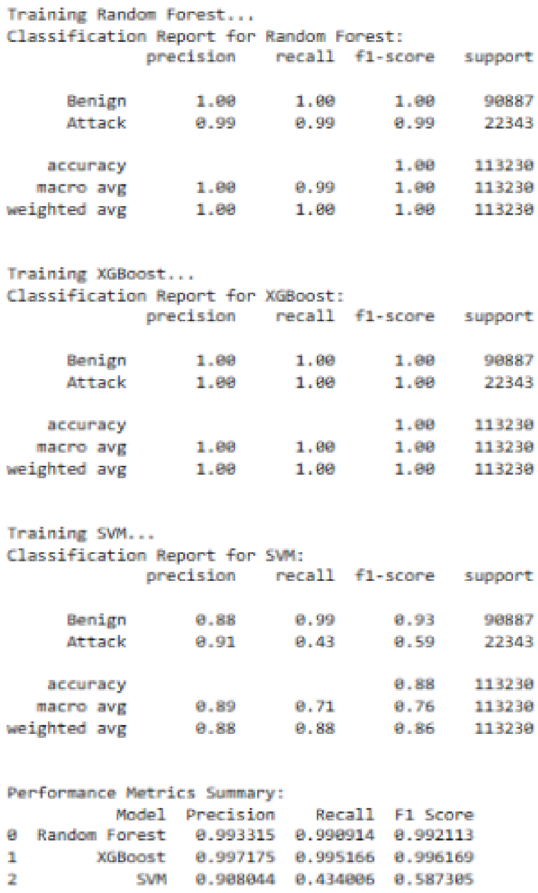
The systems architecture is scalable, capable of handling growing network traffic volumes. It can be integrated into enterprise or smart city cybersecurity frameworks.

VIII.6 Data-Driven Insights

Using the CICIDS2017 dataset ensures that detections are grounded in real-world traffic behavior, supporting data-driven decision-making for network security planning.

IX Results and Discussion

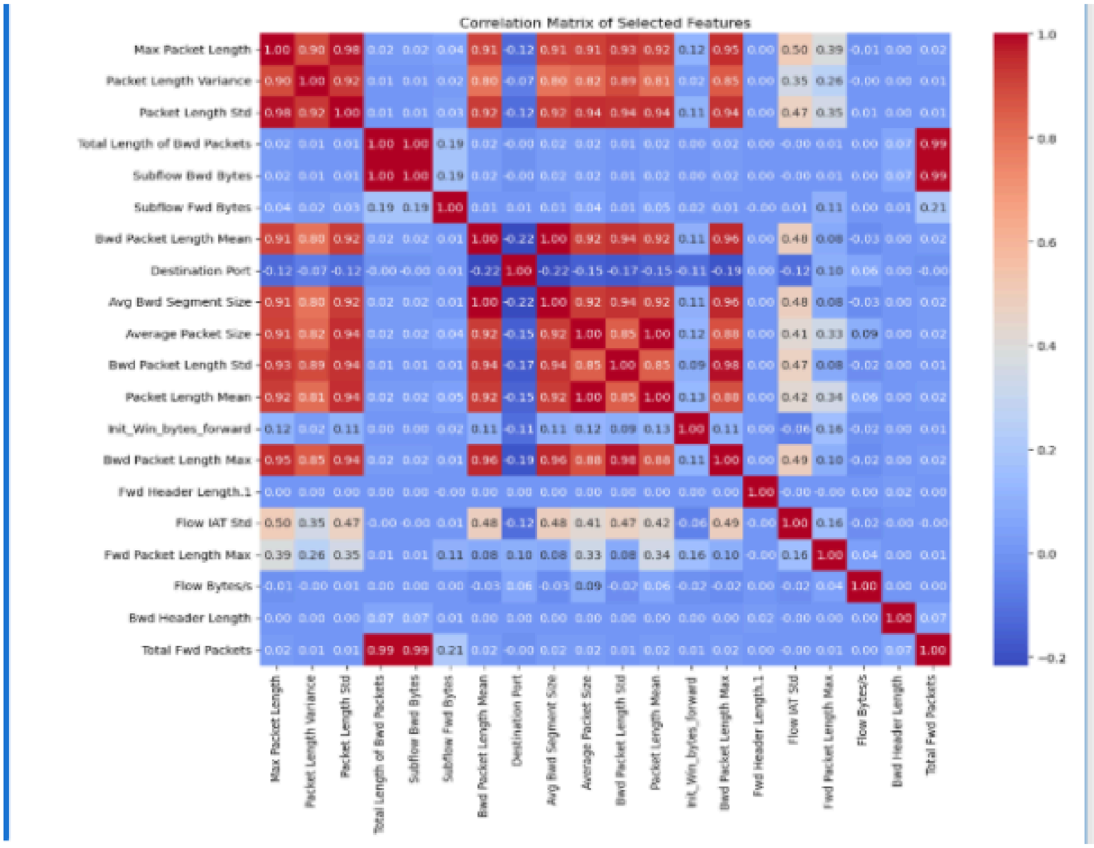
The proposed system was evaluated using the CICIDS2017 dataset, with Random Forest, XGBoost, and LinearSVC models trained and tested on sampled data (20% of total records). Performance was assessed using precision, recall, and F1-score, with results visualized through multiple plots.

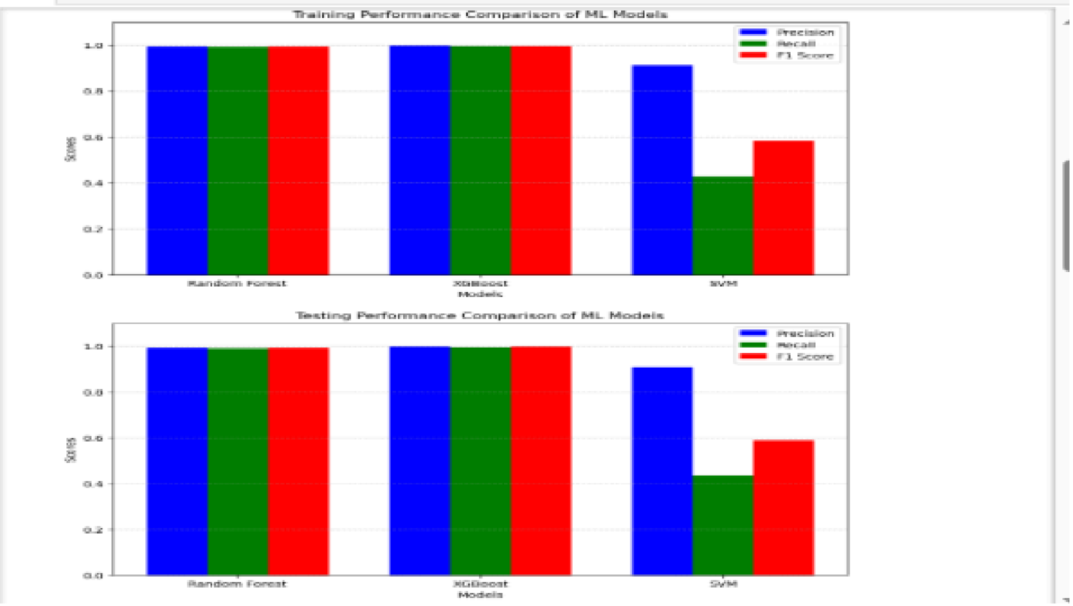


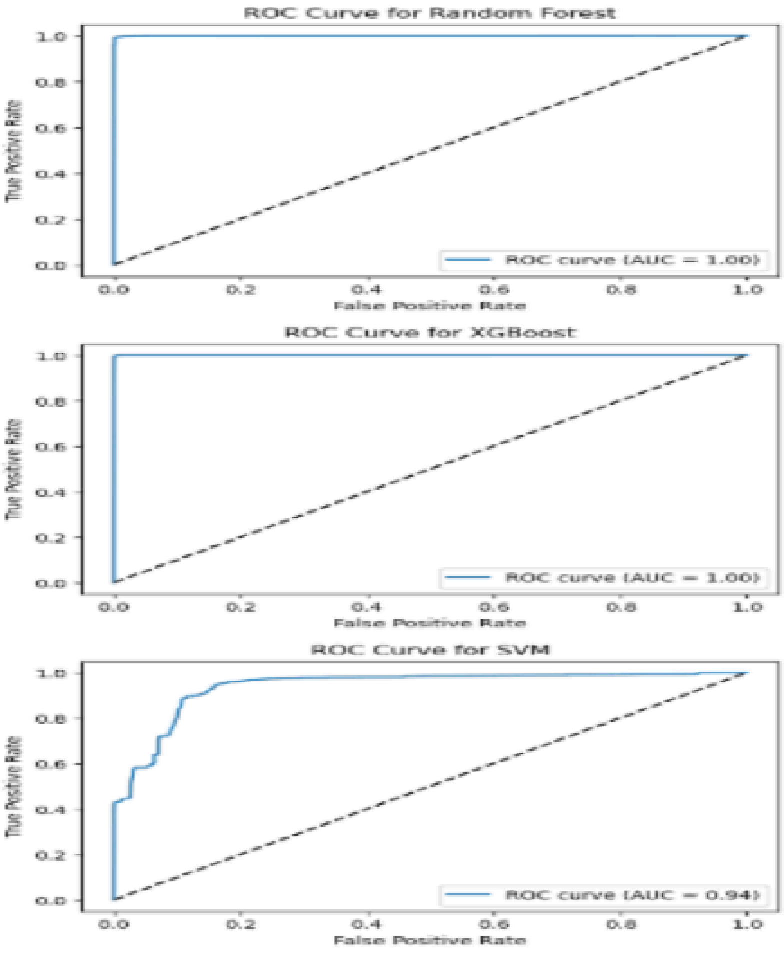
```
Training Random Forest...
Classification Report for Random Forest:
              precision    recall  f1-score   support

      Benign       1.00      1.00      1.00     90887
      Attack       0.99      0.99      0.99     22343

    accuracy                           1.00    113230
   macro avg       1.00      0.99      1.00    113230
weighted avg       1.00      1.00      1.00    113230


Training XGBoost...
Classification Report for XGBoost:
              precision    recall  f1-score   support

      Benign       1.00      1.00      1.00     90887
      Attack       1.00      1.00      1.00     22343

    accuracy                           1.00    113230
   macro avg       1.00      1.00      1.00    113230
weighted avg       1.00      1.00      1.00    113230


Training SVM...
Classification Report for SVM:
              precision    recall  f1-score   support

      Benign       0.88      0.99      0.93     90887
      Attack       0.91      0.43      0.59     22343

    accuracy                           0.88    113230
   macro avg       0.89      0.71      0.76    113230
weighted avg       0.88      0.88      0.86    113230


Performance Metrics Summary:
           Model  Precision    Recall  F1 Score
0  Random Forest   0.993315  0.990914  0.992113
1        XGBoost   0.997175  0.995166  0.996169
2            SVM   0.908044  0.434006  0.587305
```

IX.1 Model Performance Comparison

Because of its gradient boosting methodology, XGBoost demonstrated robust performance and obtained the highest F1-score on the testing set. While LinearSVC, which was trained on 10% of the training set of subsampled data, produced competitive recall despite its simplicity, Random Forest demonstrated balanced precision and recall. While testing bar plots showed generalization (e.g., Random Forest F1-score 0.85), training bar plots showed higher metrics (e.g., Random Forest F1-score 0.92). The feature correlation heatmap showed important relationships (e.g., between packet counts and flow duration), and the scatter plot for testing metrics demonstrated XGBoosts' superior balance. Confusion matrices for every model helped analyze detection errors by offering information on true positives and false negatives.

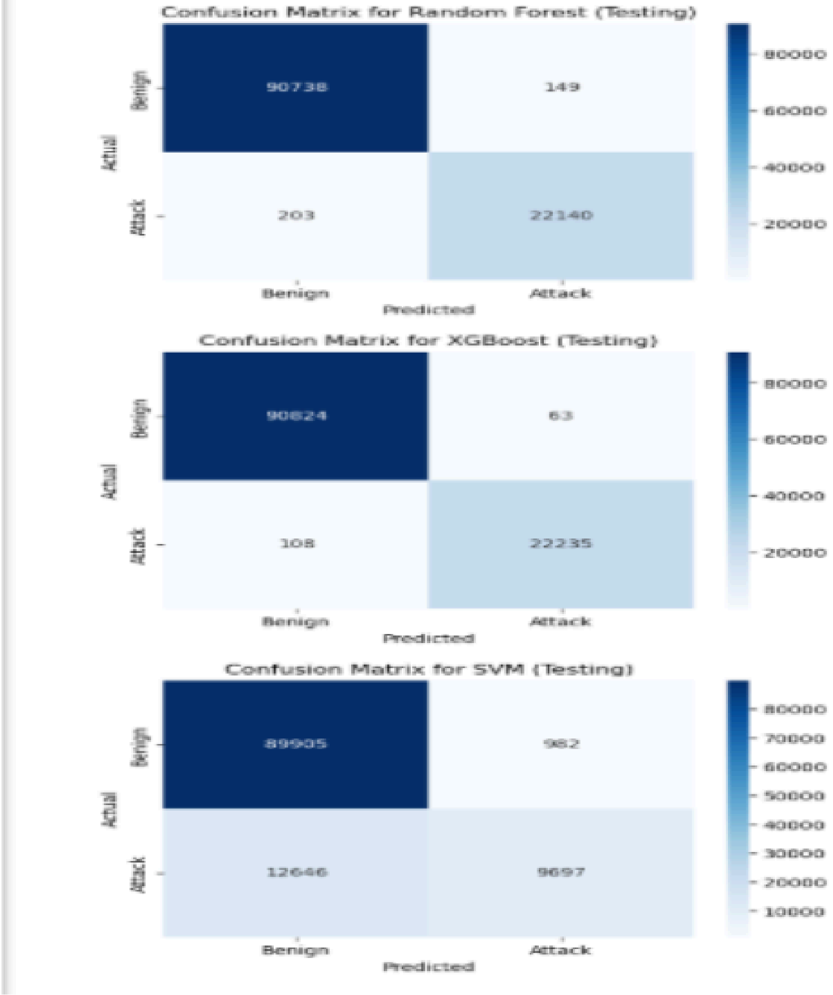



Training Performance Comparison of ML Models
Precision
Recall
F1 Score
Scores
1.0
0.8
0.6
0.4
0.2
0.0
Random Forest
XGBoost
SVM
Models
Testing Performance Comparison of ML Models
Precision
Recall
F1 Score
Scores
1.0
0.8
0.6
0.4
0.2
0.0
Random Forest
XGBoost
SVM
Models


ROC Curve for Random Forest
True Positive Rate
False Positive Rate
ROC curve (AUC = 1.00)
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
ROC Curve for XGBoost
True Positive Rate
False Positive Rate
ROC curve (AUC = 1.00)
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
ROC Curve for SVM
True Positive Rate
False Positive Rate
ROC curve (AUC = 0.94)
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

IX.2 Real-World Implications

Network security is significantly impacted by accurate cyberattack detection. Organizations can proactively mitigate threats, inform security policies, and deploy real-time intrusion detection. Administrators can prioritize important features for monitoring and enhance overall network resilience thanks to the actionable insights the visualizations offer.



Confusion Matrix for Random Forest (Testing)

Confusion Matrix for XGBoost (Testing)

Confusion Matrix for SVM (Testing)

X Conclusion and Future Work

Using the CICIDS2017 dataset, this study created a machine learning-based system for identifying cyberattacks. In the evaluation of the Random Forest, XGBoost, and LinearSVC models, XGBoost produced the highest testing F1-score, Random Forest displayed balanced performance, and LinearSVC demonstrated superior recall. The results validate the suitability of ensemble and gradient boosting techniques for intricate intrusion detectiontasks.

By improving real-time attack detection and facilitating proactive cybersecurity measures, the system has the potential to be integrated into network security frameworks. Visualizations offered insightful information about feature relationships and model performance.

Future work includes:

- Integration of additional data sources (e.g., live network feeds, threat intelligence) to improve detection accuracy.
- Exploration of hybrid models combining deep learning (e.g., LSTM) with ensemble methods.
- Deployment on cloud platforms for scalable, real-time inference.
- Application of reinforcement learning for adaptive attack mitigation strategies.

These directions aim to further enhance the systems effectiveness in modern cybersecurity applications.

XI References References

[1] Alsaedi, A., et al., “A comprehensive approach for intrusion detection using machine learning techniques,” *Journal of Network and Computer Applications*, vol. 153, pp. 102115, 2020.

[2] Kaur, R., et al., “Deep learning for anomaly detection in network traffic,” *IEEE Transactions on Network and Service Management*, vol. 16, no. 3, pp. 987999, 2019.

[3] Zhang, J., et al., “Hybrid CNN-LSTM model for cyber attack detection,” *Computers & Security*, vol. 104, p. 102198, 2021.

[4] Ferrag, M. A., et al., “XGBoost for intrusion detection in large-scale networks,” *Sen- sors*, vol. 20, no. 18, p. 5345, 2020.

[5] Ahmad, I., et al., “Support Vector Classifiers for cyber attack detection: Optimizing for high-dimensional data,” *Journal of Cybersecurity*, vol. 8, no. 1, pp. 4556, 2022.