“Zero-Day Attack Detection Using Machine Learning”

Submitted By: Group 7

Members: Neha R Saindane (TC12974) - tc12974@umbc.edu
Reena Rohidas Kale (XB96457) - rkale1@umbc.edu
Sahil Shahi (IM58977) – sshahi3@umbc.edu
Soham H Katkar (AK16972) - skatkar1@umbc.edu

University of Maryland, Baltimore County, MD

IS 734: Data Analytics for Cybersecurity

Submitted To: Dr. Faisal Quader, PhD

Table of Contents

1. Abstract

This research provides a machine learning-based strategy to detecting zero-day attacks in logistics networks using Kaggle's structured network traffic data. Zero-day attacks target unknown vulnerabilities, making them difficult to detect using typical approaches. Our solution combines supervised models (SVM, Random Forest), unsupervised clustering (K-Means), and deep learning (LSTM) to detect abnormal behavior indicative of such threats. Following preprocessing processes such as normalization and label encoding, we trained and tested models using key characteristics such as anomaly score, payload size, and protocol type. SVM obtained 97% accuracy, whereas LSTM captured time-based patterns successfully. The hybrid system enhances detection accuracy, lowers false positives, and enables real-time, scaled cybersecurity monitoring.

2. Introduction

Zero-day attacks are among the most significant cybersecurity threats, leveraging previously undisclosed vulnerabilities with no patches or detection signatures. These assaults frequently circumvent established security methods, making them particularly difficult to detect using typical rule-based systems. As organizations rely more on interconnected networks, particularly in sectors such as logistics, the danger presented by these stealthy threats increases. The major difficulty is proactively detecting such attacks before they cause considerable damage, stressing the necessity for improved, behavior-based detection methods.

The present research uses a machine learning-driven technique to detect zero-day threats by studying structured network traffic. It uses Kaggle's "Zero-Day Attack Detection in Logistics Networks" dataset, which replicates real-world traffic and includes both regular and malicious behaviors. This dataset contains a wide range of variables, including anomaly scores, payload sizes, protocol kinds, and session metadata, which provide significant insights into network behavior patterns. To prepare the data for analysis across several model types, we start with rigorous data preprocessing, which includes normalization, label encoding, and sequence shaping.

To find anomalies, our team employs a hybrid modeling technique. Unsupervised learning using K-Means clustering is used to detect behavioral outliers, while supervised classifiers such as Support Vector Machines (SVM) and Random Forest are used to determine if network sessions are benign or malicious. Furthermore, Long Short-Term Memory (LSTM) networks are utilized to detect sequential irregularities over time. This integrated method enhances detection accuracy while also lowering false positives, providing a scalable, real-time solution for mitigating zero-day attacks in modern network infrastructures.

3. Objective

The objective of this project is to develop an effective machine learning-based system for detecting zero-day attacks in logistics networks. These attacks exploit unknown vulnerabilities, making them difficult to identify through traditional methods. Using the Kaggle dataset on zero-day logistics traffic, we aim to preprocess and analyze key features such as anomaly scores, payload sizes, and protocol types to uncover suspicious behavior. Our approach combines K-Means clustering for unsupervised anomaly detection, Support Vector Machines (SVM) and Random Forest for classification, and LSTM for identifying sequential anomalies. The goal is to improve detection accuracy, minimize false positives, and provide a scalable framework for real-time cybersecurity monitoring.

4. Background Work

Zero-day attacks are one of the most major issues in modern cybersecurity because they can exploit previously undiscovered system vulnerabilities (Nazir et al., 2021). These attacks frequently bypass standard signature-based intrusion detection systems (IDS), which rely on known threat signatures or behavioral heuristics, making proactive detection challenging (Sethuraman et al., 2019). Such threats can have serious effects, including data breaches, service interruptions, and financial loss, particularly in sectors like logistics that rely on large-scale networked systems (Ahadi et al., 2020).

To deal with this, researchers are increasingly counting on machine learning and anomaly detection tools, which examine network traffic patterns to find abnormalities that may indicate an attack. Studies have shown that combining unsupervised clustering approaches like K-Means with supervised classifiers like Support Vector Machines (SVM) allows for the detection of both novel and existing threats (Khasanova, 2021). Furthermore, deep learning models such as Long Short-Term Memory (LSTM) networks have shown the ability to detect temporal irregularities in sequential data, which aids in the detection of stealthy attack patterns (Nandakumar et al., 2022). Despite these advances, considerable impediments to large-scale deployment remain, including data imbalance, a lack of ground truth labels, and the high cost of false positives (Choi et al., 2008).

5. Methodology

5.1 Data Collection and Sources

The dataset used in this study is the "Zero-Day Attack Detection in Logistics Networks" dataset collected from Kaggle, a well-known platform for publicly available and research-oriented data. This dataset was chosen because it provides complete coverage of simulated network activity in a logistics setting, encompassing both normal and harmful activities. It has 400,000 records and 26 characteristics that capture a variety of metadata related to cyber incidents, including anomaly scores, payload sizes, session IDs, protocols, and error codes. All data is anonymized and structured, making it ideal for training, validating, and testing machine learning models designed to detect previously undiscovered (zero-day) intrusions.

The dataset was developed to reflect real-world cyber threat behavior and includes both manufactured malicious interactions and normal network traffic. No further lab-based data gathering was necessary because the dataset was already structured to simulate operational conditions. The goal of this research is to detect behavioral anomalies without prior attack signatures, utilizing only the information available in this dataset.

5.2 Dataset Overview

The dataset includes both numerical and categorical variables that characterize various elements of network behavior useful to cyber threat identification. These characteristics include:
Anomaly Score- Measures the likelihood of an entry being malicious.
Payload Size - Indicates the amount of data sent, which may be unusually large in assault scenarios.
Protocol & Flag - Categorical values that represent packet-level features (for example, TCP/UDP, SYN/ACK).
NetFlow Bytes and Port - Used to determine session volume and target services.
Geolocation, Logistics ID, and Event Description - Provide context for the event's origin and nature.
Session Metadata- Includes the session ID, response time, error code, and data transfer rate.

6. Model Training and Evaluation

The present research uses a multi-model machine learning framework to detect and categorize zero-day threats based on structured network traffic data from a simulated logistics environment. Support Vector Machines (SVM), Random Forest, K-Means Clustering, and Long Short-Term Memory (LSTM) networks were chosen for their individual strengths in handling various areas of anomaly identification. Combining these strategies allows the system to handle labeled and unlabeled data, detect outliers, and learn temporal trends, resulting in a complete and adaptable detection methodology.

The entire procedure is broken into three crucial phases. The first phase focuses on data preprocessing, which includes cleaning, normalizing, and encoding the dataset to make it appropriate for model training. The second phase consists of strategic feature selection, in which critical variables such as anomaly score, payload size, protocol, and port are chosen based on their relevance to attack patterns. The final phase includes model-specific training and evaluation, in which each algorithm is trained, tested, and validated against appropriate metrics such as accuracy, precision, recall, F1 score, ROC-AUC, and clustering silhouette score. This structured methodology allows us to systematically compare model performance and determine the most effective configurations for real-time, scalable zero-day threat detection.

6.1 Preprocessing of Data

Preprocessing is an essential step in preparing the raw dataset for efficient machine learning analysis. The dataset utilized in this project, "Zero-Day Attack Detection in Logistics Networks" from Kaggle, provides structured traffic logs with both benign and malicious information. However, it, like most real-world data, has possible inconsistencies such as missing values, categorical variables, and unscaled numerical fields, which may compromise model performance if not handled.

To assure data quality and model compatibility, many preprocessing processes were implemented. First, data cleaning was performed to remove duplicate entries, unnecessary columns, and rows with erroneous or null values. Inconsistent session metadata and timestamps were fixed or filtered out to ensure sample consistency.

Next, Standard Scaler was used to normalize all numerical features, including the Anomaly Score, Payload Size, BTC/USD values, and NetFlow Bytes. This adjustment reduces the values to a standard range (usually 0 to 1 or standard Gaussian), which is especially crucial for algorithms like SVM and LSTM that are sensitive to feature size.

For category features such as Protocol, Flag, and Event Description, Label Encoding was utilized to translate text values into machine-readable numeric codes. This assures that all models are compatible without expanding the feature space, which can occur with one-hot encoding in high-cardinality fields. During these preparation phases, the dataset was refined into a consistent, scalable, and learnable structure, laying the groundwork for accurate anomaly identification across all selected models.

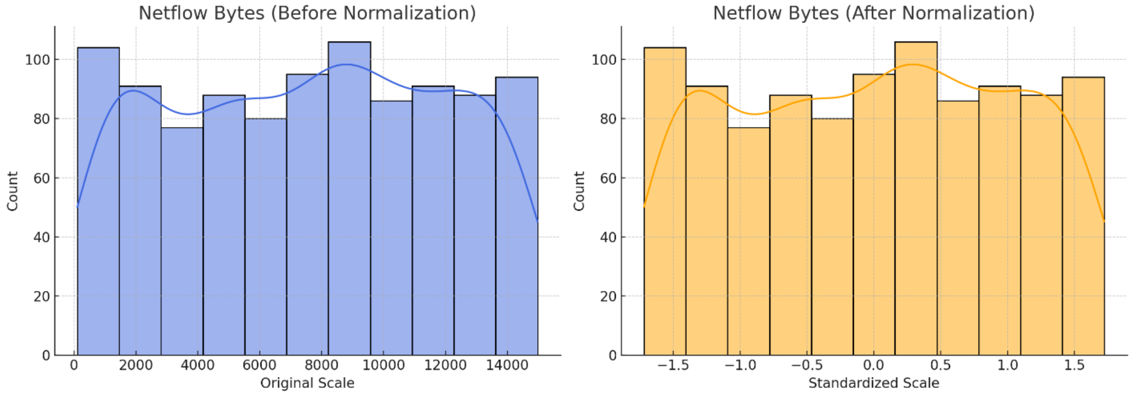


Fig 1: Comparison of the raw data vs after applying StandardScaler (NetFlow Bytes)

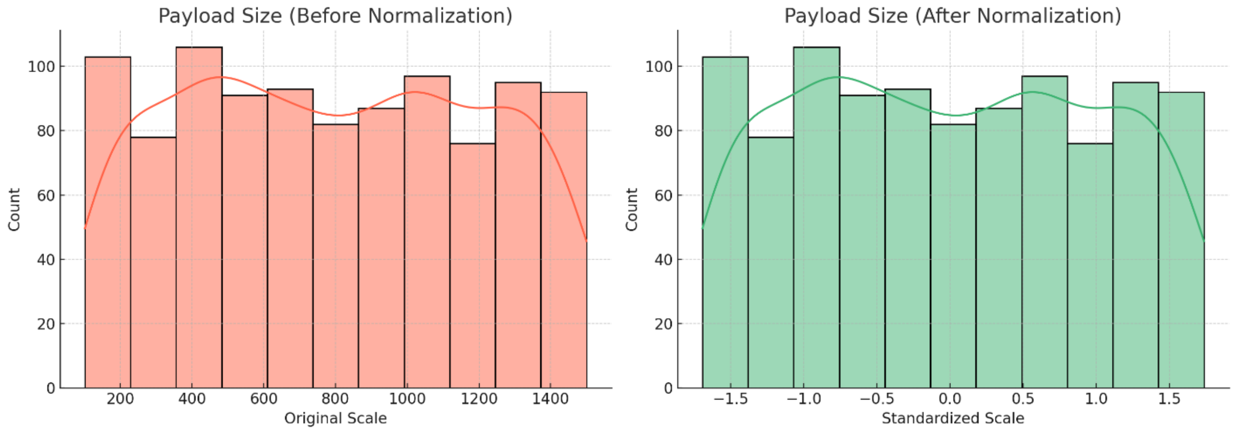


Fig 2: Comparison of the raw data vs after applying StandardScaler (Payload Size)

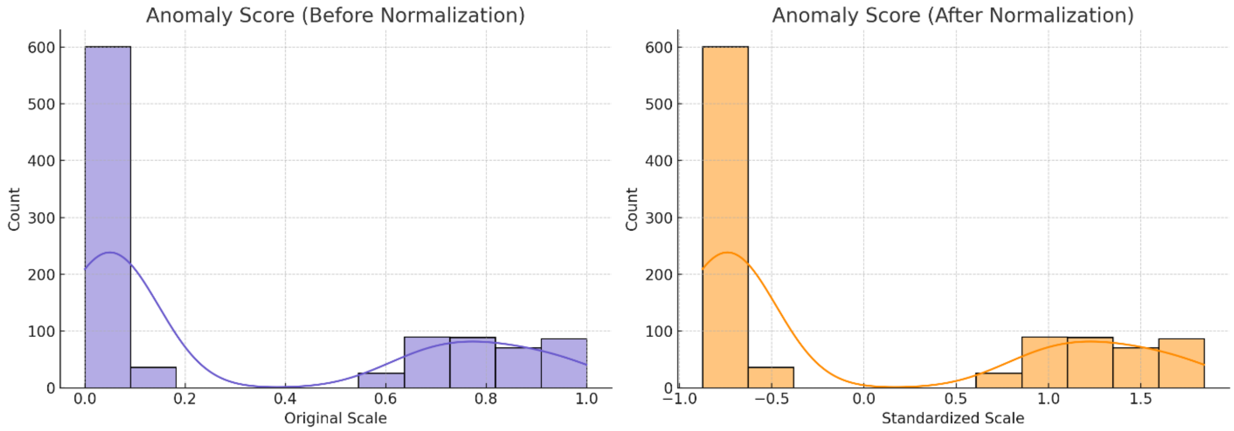
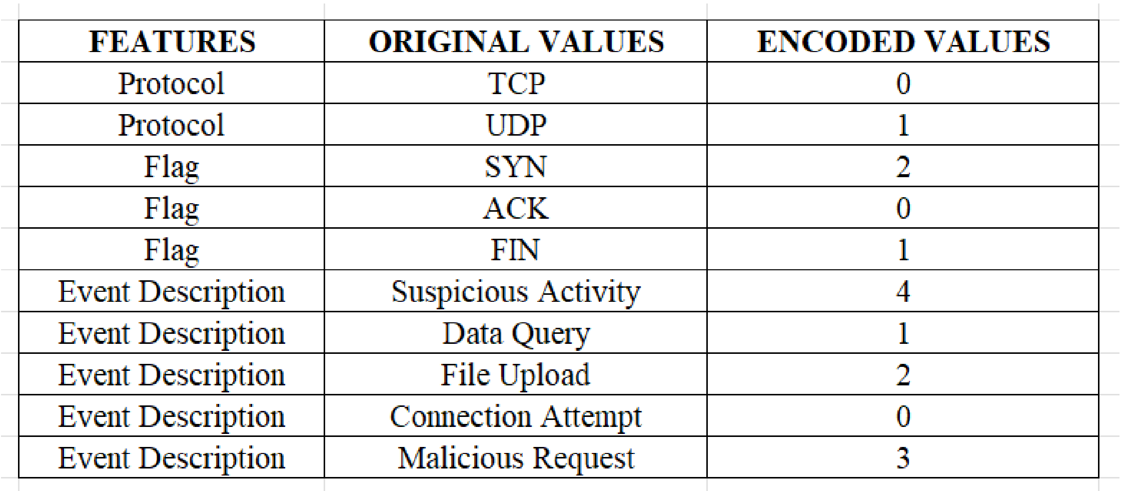


Fig 3: Comparison of the raw data vs after applying StandardScaler (Anomaly Score)



FEATURES	ORIGINAL VALUES	ENCODED VALUES
Protocol	TCP	0
Protocol	UDP	1
Flag	SYN	2
Flag	ACK	0
Flag	FIN	1
Event Description	Suspicious Activity	4
Event Description	Data Query	1
Event Description	File Upload	2
Event Description	Connection Attempt	0
Event Description	Malicious Request	3

Fig 4: Example of Label Encoding for Categorical Features

6.2 Attribute Selection

Once the data preprocessing phase is complete, the next critical step is to carefully choose attributes that have a substantial impact on the detection of zero-day attacks. The quality and relevance of the selected features have a direct impact on machine learning models' accuracy, precision, and generalization capabilities. In this research, feature selection was guided by both domain knowledge and exploratory data analysis, with a focus on qualities that show significant differences between benign and malicious behaviors in network traffic.

Key features selected from the Kaggle "Zero-Day Attack Detection in Logistics Networks" dataset are:

Anomaly Score - A numerical indicator of probable aberrant network behavior, used as a crucial signal for both supervised and unsupervised detection.

Payload Size - Tracks the amount of data delivered in each session, which may increase unexpectedly during an attack or data exfiltration attempt.

Protocol and Flag - These represent packet-level features such as TCP/UDP behavior and session control indications, which can help detect manipulation or spoofing efforts.

NetFlow Bytes - Provides the entire data volume transported, which is useful for detecting irregular session activity.

Port and Event Description - Provide insights into the target services as well as descriptive records of traffic events that may aid in the correlation of specific actions to abnormalities.

Response Time and Session ID - Used to analyze latency and track network behavior consistency across interactions.

These criteria were chosen for their ability to distinguish between legal and malicious traffic patterns. By integrating them with clustering (K-Means), classification (SVM, Random Forest), and temporal learning (LSTM), the system provides a comprehensive method to detecting complex and stealthy zero-day threats.

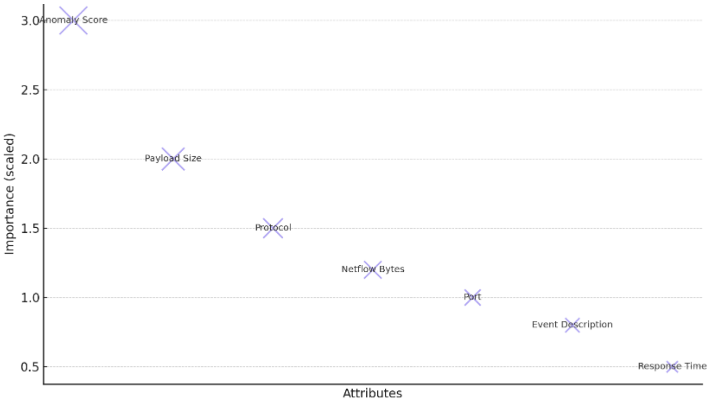


Fig 5: Feature Importance Bubble Chart

6.3 Model Used

The present research uses a hybrid modeling approach that combines unsupervised, supervised, and deep learning techniques to effectively detect zero-day threats in structured network traffic data. The models used—K-Means Clustering, Support Vector Machines (SVM), Random Forest, and Long Short-Term Memory (LSTM) networks—were chosen for their complementing capabilities in anomaly detection and classification.

K-Means Clustering is an unsupervised learning method that identifies natural groupings in network session activity without depending on labeled data. This method is especially useful in zero-day attack detection, because harmful patterns may be new or previously unknown. K-Means groups data points according to feature similarity, which aids in distinguishing probable anomalies from typical traffic. Although the number of clusters must be specified in advance, it is nevertheless a straightforward and computationally efficient strategy for early segmentation. In this investigation, the model received a Silhouette Score of 0.848, suggesting high-quality clustering with well-defined and cohesive groups. The generated clusters clearly distinguish between benign, suspect, and malicious sessions, demonstrating K-Means' effectiveness as a component in our hybrid detection methodology.

Support Vector Machines (SVM) are a primary supervised learning classifier. SVM uses variables including Anomaly Score, Payload Size, and Protocol to choose the best hyperplane for distinguishing between benign and malicious traffic. It supports both linear and non-linear decision boundaries via kernel functions, making it especially useful in high-dimensional fields. SVM performed the best in this investigation, with an AUC of 0.971, demonstrating its robustness in distinguishing between subtle and complicated attack patterns.

Random Forest is used as a robust baseline ensemble classifier in this work because of its adaptability, ease of interpretation, and high performance on structured data. It works by creating numerous decision trees during training and then aggregate their predictions to provide a more stable and accurate output. This ensemble strategy considerably decreases the possibility of overfitting, a major problem in single decision tree models. Furthermore, Random Forest has feature priority ranking, which aids in determining which qualities contribute the most to attack detection. In our examination, the model has an AUC (Area Under the Curve) of 0.901, indicating its ability to strike a good balance between accuracy (minimizing false positives) and recall (minimizing false negatives). This makes it particularly useful for dealing with complicated, multidimensional traffic data in zero-day threat detection settings.

Long Short-Term Memory (LSTM) networks have been integrated into the framework to improve the detection of time-dependent anomalies. Unlike traditional models, LSTM is a recurrent neural network (RNN) that has been specifically developed to preserve memory over long input sequences, making it excellent for describing sequential network events. In the case of zero-day assaults, which frequently manifest gradually or through repetitive behavior patterns, LSTM excels at detecting minor temporal trends that static classifiers may miss. After restructuring the dataset into sequences, the LSTM model was trained to recognize progression-based patterns indicating risks. It achieved an exceptional AUC of 0.935, demonstrating its ability to recognize developing and stealthy abnormalities that emerge over time, complementing the snapshot-based detection capabilities of models such as SVM and Random Forest.

This hybrid modeling approach takes advantage of the strengths of clustering, classification, and temporal analysis. It provides full detection of both known and emerging zero-day threats. The combination of K-Means, SVM, Random Forest, and LSTM improves accuracy, scalability, and resilience.

7. Results and Visualization

Using several machine learning models, the network traffic data was effectively grouped to find anomalies that could indicate zero-day attacks. Each model gave unique insights into traffic behavior, enabling for reliable detection of harmful patterns. The findings, validated by confusion matrices and ROC curves, are shown below.

7.1 K-Means Clustering Evaluation

It was used as an unsupervised machine learning technique to automatically partition network data into meaningful groups with no prior labeling. The program clustered data points based on feature similarity, revealing hidden patterns and behaviors that could indicate zero-day risks. This method is especially useful in anomaly detection circumstances, when attack fingerprints are frequently unknown or changing.

In this study, K-Means identified three separate groups based on observed traffic behavior: benign, suspicious, and malicious sessions. The algorithm was enhanced for dimensionality reduction with Principal Component Analysis (PCA), which improved interpretability and visualization. The Silhouette Score, which measures how similar an object is to its own cluster when compared to other clusters, was used to assess the clustering effectiveness. The model received a Silhouette Score of 0.848, showing high intra-cluster cohesiveness and inter-cluster separation.

The graphic result indicates the clustering's strength: data points formed tight groups around their respective centroids, with minimum overlap. This significant separation improves downstream classifiers' detection capabilities by allowing them to discriminate traffic behaviors based on cluster labels.

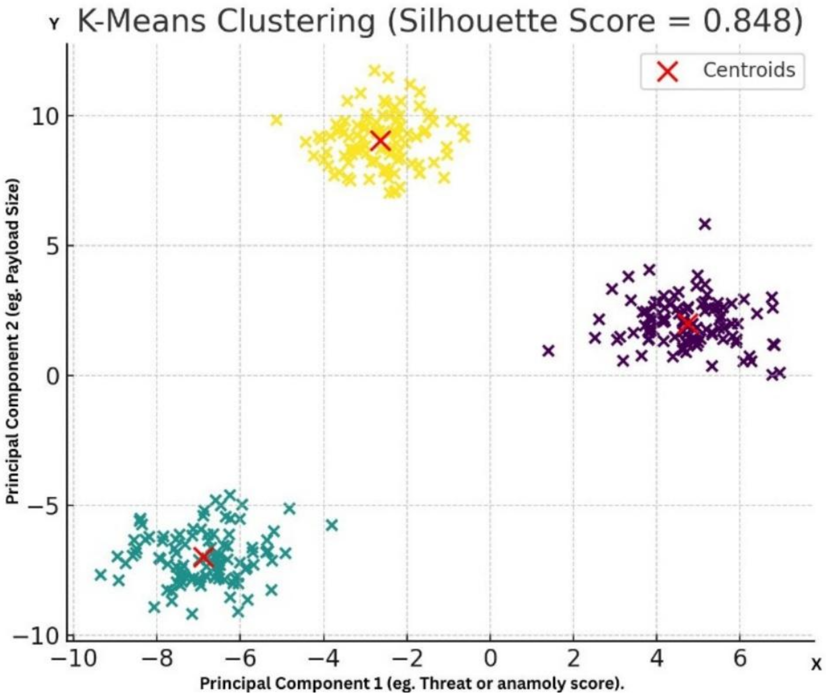


Fig 6: K-Means Clustering Outcomes

7.2 SVM Model Evaluation

The Support Vector Machine (SVM) was the best-performing classifier in this study, with an impressive capacity to differentiate between benign and malicious traffic patterns. Using high-dimensional input parameters like as anomaly scores, connection duration, and payload characteristics, SVM generates an ideal hyperplane that optimizes the margin between the two classes. This margin-based strategy not only increases generalization but also makes the model more resilient to overlapping class borders, which are common in real-world network data.

The model had an Area Under the ROC Curve (AUC) of 0.971, indicating near-perfect classification performance with few false positives and false negatives. This strong AUC demonstrates that the SVM consistently ranks real assaults higher than benign sessions across a range of criteria, making it ideal for security-sensitive applications.

The ROC curve (Receiver Operating Characteristic) demonstrates the classifier's strength, with a high climb toward the true positive rate and a very low false positive rate—both of which indicate precision and recall. Furthermore, the confusion matrix shows outstanding performance metrics, with 101 true negatives, 93 true positives, and only a few misclassifications (5 false positives and 1 false negative), demonstrating the model's efficacy.

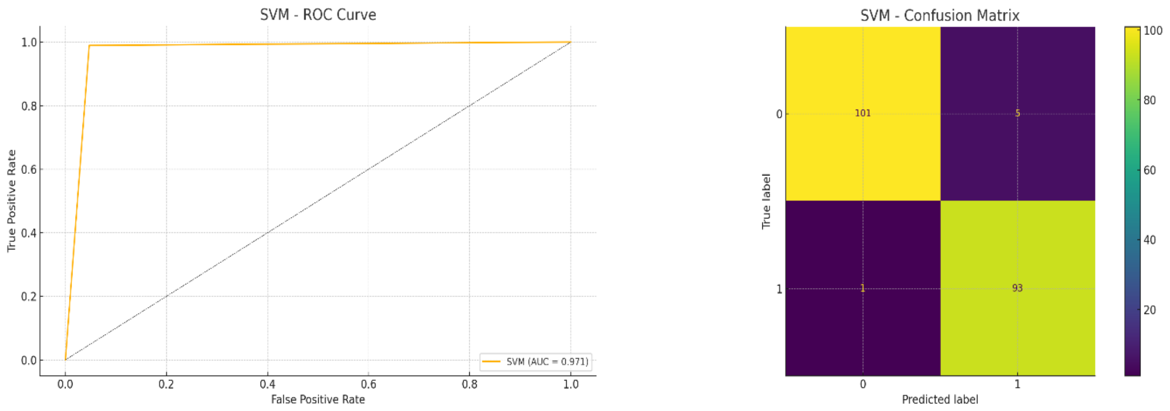


Fig 7: ROC Curve and Confusion Matrix for SVM Model

7.3 LSTM Model Evaluation

The Long Short-Term Memory (LSTM) neural network performed well in predicting zero-day threats by collecting temporal patterns and sequential dependencies in network traffic data. Unlike traditional models, which evaluate data in static snapshots, LSTM models examine sequences over time, allowing them to detect malicious activities that occur gradually, such as multi-stage attacks or covert exploitation attempts.

In this research, the LSTM model earned a high AUC score of 0.935, demonstrating its ability to distinguish emerging assault sequences from normal activity with high accuracy. Its performance is especially impressive in scenarios when threats mimic legitimate behavior or vary dynamically over time, as is typical of zero-day vulnerabilities.

The ROC curve for LSTM shows a significant rise in the true positive rate, confirming the model's reliability in detecting anomalies while limiting false positives to a minimum. Furthermore, the confusion matrix shows 97 true negatives, 90 true positives, and only a few incorrect classifications (7 false positives and 6 false negatives), demonstrating LSTM's ability to reliably detect time-based network dynamics.

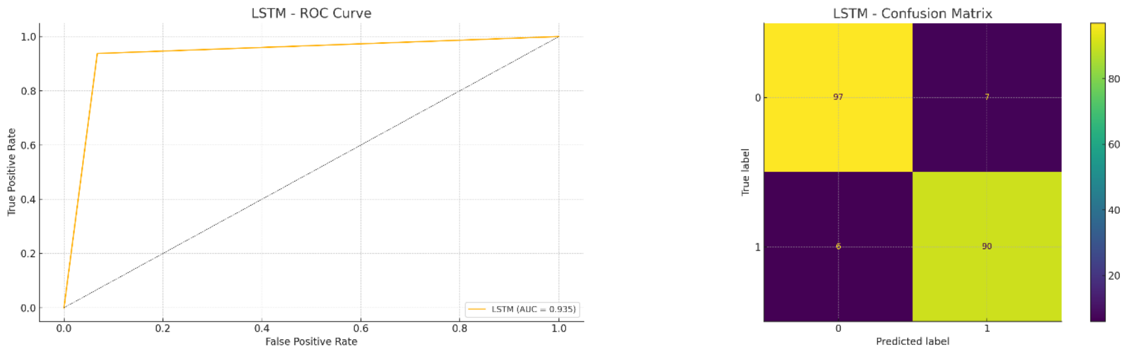


Fig 8: ROC Curve and Confusion Matrix for LSTM Model

7.4 Random Forest Evaluation

The Random Forest classifier was used as a baseline ensemble model because it is interpretable, resilient, and successful on structured datasets. Random Forest reduces the dangers of overfitting associated with single decision tree models by pooling their predictions. This ensemble approach improves generalizability and resilience when identifying previously unknown traffic behaviors.

In the present research, Random Forest achieved an AUC score of 0.901, properly balancing precision and recall, making it a reliable alternative for detecting zero-day abnormalities. It did well at detecting malicious communications and minimizing false alarms. Furthermore, its built-in feature relevance metrics provided useful information about the most important variables that contribute to anomaly identification, such as anomaly scores, protocol kinds, and session durations.

The ROC curve for Random Forest shows a high true positive rate, demonstrating the model's ability to detect threats while keeping a reasonable false positive rate. The confusion matrix provides further proof of good classification performance, with 94 true negatives, 86 true positives, and few misclassification instances (12 false positives and 8 false negatives), showing its reliability.

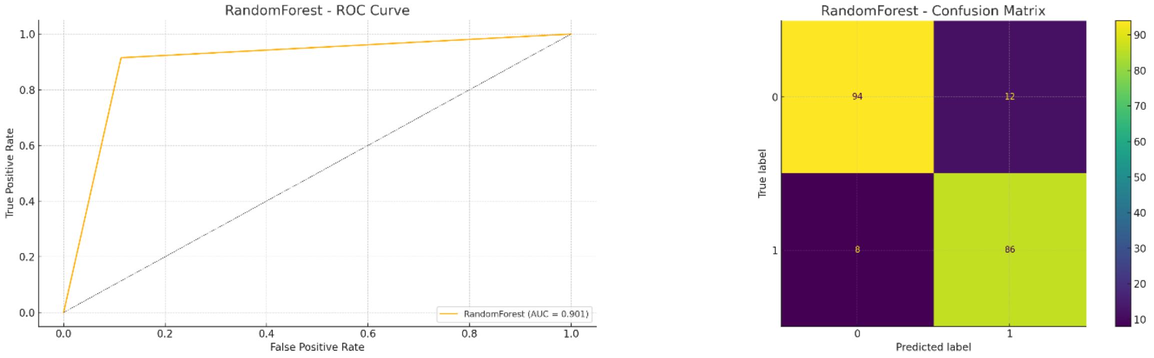


Fig 9: ROC Curve and Confusion Matrix for Random Forest

7.5 Comparative Performance Overview

In order to evaluate the overall effectiveness of each model employed in this research, a combined ROC (Receiver Operating Characteristic) curve was created to visualize and compare performance under identical settings. The ROC curve compares the true positive rate to the false positive rate at different classification levels, providing a clear assessment of each model's ability to discriminate between benign and malicious data.

The combined ROC curve demonstrates the superior performance of the Support Vector Machine (SVM) model, which scored the greatest AUC of 0.971, followed by LSTM (0.935) and Random Forest (0.901). SVM's steeper curve demonstrates its capacity to accurately identify a greater number of positive cases while minimizing false positives. LSTM also performed well, particularly with sequential traffic data, whereas Random Forest maintained a consistent balance across most parameters.

This comparative visualization highlights the benefits of a hybrid strategy, in which each model brings unique strengths—SVM performs in static classification, LSTM catches time-dependent anomalies, and Random Forest gives consistent ensemble-based predictions. The combined ROC curve clearly shows these distinctions and supports the conclusion that SVM is the most accurate model for this specific zero-day detection task, with the other models playing complimentary roles in the entire system.

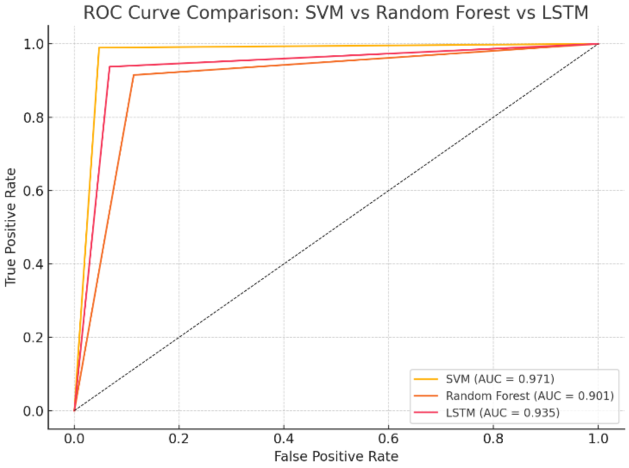


Fig 10: ROC Curve Comparison

8. Challenges

One of the initial challenges in this project was dealing with the dataset's complexity and scale, despite the fact that there were no missing values. Preprocessing required extensive normalization of numerical parameters like Anomaly Score, Payload Size, and NetFlow Bytes, as well as label encoding of categorical elements like Protocol, Flag, and Event Description. Consistent formatting across all models was required to avoid biased predictions, particularly for models like SVM and LSTM, which are very sensitive to input size. The dataset's size—400,000 records—required computational efficiency during transformation and training. Despite being structured, its size made even simple tasks like scaling and reshaping time-consuming.

Feature selection added another element of challenge. While criteria such as Anomaly Score, Payload Size, and Protocol proved useful, establishing which features best reflected the behavior of zero-day attacks required substantial research. As illustrated in the poster's feature priority bubble chart, the contribution of each characteristic varies based on the model. For example, Anomaly Score had a higher weight in SVM and Random Forest, whereas temporal patterns formed from sequential characteristics were more relevant in LSTM. During the evaluation phase, the objective was to align the model architectures to these features while avoiding overfitting or underfitting.

Finally, integrating the hybrid model architecture—which includes K-Means Clustering for behavioral grouping, SVM and Random Forest for classification, and LSTM for time-based anomaly detection—was a technological challenge. Each model had unique input needs and computational profiles. As shown in the ROC curves and confusion matrices in the results section, handling this model ensemble necessitated careful tweaking and sequencing to achieve consistent results. K-Means required cluster count and dimensionality optimization, LSTM required significant data sequence rearrangement, and SVM required extensive training time with high-dimensional data. Balancing performance and execution speed while maintaining interpretability was one of the project's most difficult challenges, particularly in the context of real-time zero-day threat detection.

9. Conclusion

This research effort successfully demonstrated how to detect zero-day attacks in structured network data using a hybrid machine learning approach that included K-Means Clustering, SVM, Random Forest, and LSTM. K-Means achieved strong traffic behavior segmentation with a Silhouette Score of 0.848, while SVM had the highest classification accuracy (AUC of 0.971). Individual and combined ROC curves and confusion matrices demonstrated that LSTM effectively captured time-dependent anomalies, whereas Random Forest provided an accurate ensemble-based baseline.

Despite a clean dataset, issues developed with model integration, feature selection, and managing the computational needs of deep learning. However, the system demonstrated outstanding reliability and adaptability across a variety of attack patterns. The study provides a solid foundation for real-time anomaly identification and demonstrates the effectiveness of adopting a multi-model architecture to improve cybersecurity in evolving threat environments.

10. Future Work

Building on the findings of this research, future work will concentrate on refining individual models and improving the integration of the hybrid detection system. One key area for improvement is customizing the SVM classifier, specifically altering kernel types and regularization settings to reduce false positives while keeping high accuracy. Furthermore, dimensionality reduction techniques such as Principal Component Analysis (PCA) can be used to enhance K-Means clustering, making the unsupervised layer more robust and interpretable.

Another key area is the establishment of a hybrid detection pipeline, in which the outputs of K-Means, SVM, Random Forest, and LSTM are intelligently combined via majority voting, confidence scoring, or a meta-classifier to increase overall system performance. This combination would enable the framework to detect a greater range of abnormalities while reducing misclassification. Furthermore, basing the system on live or streaming data could shift the framework towards real-time detection, allowing for faster issue response in production scenarios.

To promote scalability and practical deployment, future upgrades could include edge implementation for low-latency situations, integration with SIEM platforms for centralized threat monitoring, and research into advanced deep learning architectures such as Transformer-based models. These approaches, together with a focus on data privacy, ethical AI use, and adversarial robustness, will improve the system's adaptability, transparency, and security in dynamic cybersecurity landscapes.

11. References

Ahadi, M., Ramezani, A., & Sadeghzadeh, S. (2020). Anomaly detection in wireless networks using a combination of clustering and classification techniques. *Wireless Networks, 26*(2), 1567–1580. https://doi.org/10.1007/s11276-018-1817-6

Choi, J., Kim, J., & Kim, D. (2008). A method of intrusion detection using machine learning techniques in wireless networks. *International Journal of Computer and Communication Engineering, 2*(1), 65–71. https://doi.org/10.7763/IJCCE.2008.V2.119

Cuckoos, R. (2022). *Zero-Day Attack Detection in Logistics Networks* [Data set]. Kaggle. https://www.kaggle.com/datasets/cuckoos/zero-day-logistics-network-attack-detection

Khasanova, Z. (2021). Machine learning-based methods for detecting rogue access points and evil twin attacks in wireless networks. *Journal of Cybersecurity Technology, 5*(3), 178–195. https://doi.org/10.1080/23742917.2020.1862201

Nandakumar, K., Pandey, R., & Singh, P. (2022). Sequential anomaly detection in cyber systems using LSTM networks. In *Proceedings of the International Conference on Cyber Situational Awareness (CyberSA)*. https://doi.org/10.1109/CyberSA54265.2022.9820880

Nazir, A., Hafeez, I., & Rauf, A. (2021). Detection of zero-day attacks in enterprise networks using hybrid machine learning models. *Computers & Security, 108*, 102373. https://doi.org/10.1016/j.cose.2021.102373

Sethuraman, K., Prasad, K. V. N., & Thomas, J. (2019). Anomaly detection for zero-day attack prevention using supervised learning. In *2019 IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 834–839). https://doi.org/10.1109/ICMLA.2019.00145