Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 12, 2025

A COMPARATIVE EVALUATION OF DIFFERENT MACHINE LEARNING TECHNIQUES IN MALWARE DETECTION & ANALYSIS

Amna Saeed Kharal[1], Ali Mukhtar[2], Arshee Ahmed[3], Muhammad Zulkifl Hasan[4], Muhammad Zunnurain Hussain[5*]

[1]*Student, Department of Computer Science, FAST-NU, Lahore, Pakistan*
[2]*Student, Department of Computer Science, FAST-NU, Lahore, Pakistan.*
[3]*Assistant Professor, Iqra University Pakistan & Multimedia University Malaysia*
[4]*Faculty of Information Technology, Department of Computer Sciences, University of Central Punjab, Lahore Pakistan*
[5]*Department of Computer Sciences, Bahria University Lahore Campus, Lahore Pakistan*

[1]l191225@lhr.nu.edu.pk, [2]l200967@lhr.nu.edu.pk, [3]dr.arshee@iqra.edu.pk, [4]zulkifl.hasan@ucp.edu.pk, [*5]zunnurain.bulc@bahria.edu.pk

DOI: https://doi.org/10.5281/zenodo.17828177

Keywords
Malware, Machine Learning, Cybersecurity, Malware Detection, Feature Selection, Random Forest Classifier, LSTM Networks, Precision and Recall

Article History
Received on 10 October 2025
Accepted on 23 November 2025
Published on 05 December 2025

Corresponding Author: * Muhammad Zunnurain Hussain

Abstract

The persistent evolution of malware and its growing sophistication poses a significant challenge to cybersecurity. This study undertakes a comparative analysis of several machine learning approaches—namely, clustering-based anomaly detection, supervised learning using Random Forest classifiers, and time-series analysis via LSTM networks—for the purpose of malware detection. Using a dataset of over 40,000 records with 25 features, the study evaluates feature extraction, scaling, and performance metrics including precision, recall, and F1-score. Our findings highlight the superior performance of security-related features in classification tasks and the necessity of fine-tuned LSTM models for time-dependent intrusion detection. The comparative insights aim to aid cybersecurity professionals in selecting optimal machine learning strategies for robust malware detection.

INTRODUCTION

The continuous proliferation and sophistication of malware have rendered traditional detection mechanisms increasingly insufficient. To address this challenge, machine learning (ML) has emerged as a powerful paradigm, offering dynamic and adaptive capabilities for threat identification and mitigation. This paper focuses on evaluating multiple ML-based strategies for detecting malware in network traffic data, emphasizing the impact of feature selection and model tuning on detection performance. By incorporating supervised, unsupervised, and time-series models, the study provides a comprehensive perspective on leveraging ML to enhance cybersecurity infrastructure.

2. Related Work

This discovery laid a foundation for investigating machine learning ways to reinforce and advance its malware endpoint detection methods and mechanisms. In their work, Vinayakumar et al. (2019) proposed ScaledMalNet which was a scalable framework leveraging the power of deep learning architecture and image processing to detect and categorize malwares[^2]. This creative solution was

applied to the MLAs to determine that deep learning algorithms were more powerful compared with the MLAs, and the solution could be used against the challenges brought in by the continuously evolving malware threats. Hajraoui and El Merabet (2019) argued about machine learning methods that are used in malware detection and the importance of features selection and classification techniques [^3]. Survey of ML approaches displayed the difference in their performances in detecting malware which in turn reveals the diversity of ML techniques. Chintha et al.(2020) insisted that machine learning tends to generally automate malware detection and analysis, putting forward several methods that have proven strongly functional in malware filtration[^4]. This comparison will direct researchers to take advantage of complex algorithms and thus, not only the malware is detected but also the machine is more secured. Moubrack and Feghali (2020) examined artificial intentionally, by malware, where the random forest classifier is found to be an effective detector[^5]. Their research only highlights the real challenge that malware developers all over the world are facing every day as cybersecurity defenses are improving. Machuche et al. (2020) meticulously investigated traditional machine learning algorithms for malware detection specifically covering deep learning and hybrid techniques[^6]. This research undertaking generates useful understanding of the evolutions and trends of malware detection technology, which assists in the assessment the performance development of malware detection technology. Therefore, images visualization-based malware detection method proposed by Sheneamer et al. (2022), which has proved to be accurate. Please find the footnote reference at the end of the text. Through an empirical standpoint, this study points at the capacity visual, side by side analysis of data might have on attaining accurate malware detection.

What Alqahtani (2021) focused on was the use of some machine learning methods for the malware detection concluding that with those techniques high accuracies can be achieved. Also cybersecurity research is becoming multidisciplinary nature.

It presents the main difficulties and the development prospects of building multi-faceted and highly complex data sources for real estate application purposes. Rkhouya and Chougdali (2021) paved the way for the Random Forest algorithm detecting malware, successfully confirming it in a big dataset. Nevertheless, their results are in harmony with the earlier studies that proved the algorithm's effectiveness. "The authors(Odintsov et al., 2022) introduced polymorphic malware, proving higher accuracy in the detection of DT, CNN, and SVM," - The authors[^10] emphasized on the challenge of polymorphic malware demonstrating high accuracy in the detection of DT, CNN, and SVM. This paper illustrates the reality of the modern malware attack and the role of expanded antimalware techniques.

These collaborative efforts result in the larger body of cybersecurity literature, with a subarea in particular being in the domain of the machine learning technique enabled to fight malware off from computers. The comparison of different machine learning solutions including supervised, unsupervised and clustering techniques in these projects gives us a practical place to realize our research which is identifying the most effective machine learning technique for intrusion detection.

3. Methodology

My research has undertaken a strategic and thorough method in the area of the effect of various machine learning methods in classifying malware from a data set comprising 40,000 items and 25 features. The following steps were taken. The following steps were taken:

1. Dataset Loading

The process is started off by having the dataset being loaded into our analytical environment for faster access and fast memory usage to enhance the processing capabilities.

2. Data Preprocessing

Our specific processing procedures were applied to clean the dataset of biases and unevenness, thus assuring the reliability of our findings in the analysis.

3. The EDA(Exploratory Data Analysis) process is crucial for the initial assessment of the dataset

before further modeling takes place. Next, we proved the data preprocessing phase and then set out to conduct an exploratory data analysis aimed at discovering hidden patterns and understanding traits of the sample. While the EDA is a multi-layered tool, it is equipped with several contributing analyses that are designed to target various elements of network

traffic and cyber threats. Graphs complement each component investigated and allow a visual comprehension of what takes place as per the considered data. Protocol Distribution: A network traffic analysis report is incomplete without a graph showing the frequencies of different network protocols used by the operators for connecting to the network, as it helps understand the traffic types and their vulnerability to a cyber attack.

Attack Type Distribution: Purposeful segregation of network traffic events by attack type (such as DDoS, malware, phishing) is vital in identification of prevalent threats and itemization of areas that require heightened security measures. Traffic Type Distribution: Identifies traffic types that are within the centre scope of an observer by name like- HTTP and DNS to understand which communications is mostly being hijacked or is compromised. Action Taken Distribution: It demonstrates actions towards detection events (for example statistics on the numbers of the unsafe drivers blocked or logged), which are the key indicator of evaluation the effectiveness of the traffic enforcement system and its policies. Severity Level Distribution: Contributes severity levels per event, which is useful for risk prioritization and incident classification. Packet Length Distribution: A histogram with KDE showing packet lengths, providing insights into typical and atypical sizes that might indicate normal or malicious activities. Source Port Distribution: Reveals common and unusual source port patterns that might suggest malicious activities through a histogram with KDE.

Anomaly Scores Distribution: Another histogram with KDE detailing anomaly scores distribution, crucial for tuning detection algorithms and threshold settings.

4. Feature Extraction

Following data preprocessing and exploratory data analysis, we proceed to extract relevant features from the dataset that are crucial for effective malware detection using machine learning techniques. Some categories of features that were focused are given as below.

A. Categorical Features

These are the features that were directly present in the dataset and did not require preprocessing, some of them are:

- IP Addresses: Transformed into a numerical format to facilitate analysis. For example, an IP address "192.168.1.1" can be converted into an integer like 3232235777. This is crucial as it allows mathematical operations and model processing.
- Ports: Source and destination ports can indicate the type of service or applications being used. They are directly used as features as they can sometimes suggest normal or malicious traffic based on known port usages.
- Packet Lengths: Directly used to understand the size of the packets being transmitted.Statistical information about packet sizes can help in identifying anomalies (e.g., very small or large packets that deviate from typical patterns).
- Timestamps: Converted into multiple features like hour of the day, day of the week, and minutes which can help identify patterns or attacks based on timing.

B. Statistical Features

- We narrow our statistics down to computing related metrics over specified time windows of packets aggregates. These statistics are in fact, working as a guiding tool for us to form opinions about these traffic patterns during the timeframe stated before.
- Mean, Variance, and Standard Deviation: We can carry out that kind of statistics for constructive decisions and anomalies can be detected by outliers in the packet length and interval characterization of typical traffic flows. The deviations from these norms may be an indication of the malicious host that may be hiding and operating within the network.
- Sum and Count: Through a method involving addition of number of packets, and data volume in the window, we can notice activities burst within this time span. The sharp peaks show up as DoS (Denial of Service) attacks and data exfiltration, an

action where large amounts of data openly leave the network.
- They represent the foundation of the statistical analysis of the network, with primary measures of network health, and thereby find recurring patterns that point to the presence of the malware.

C. Flow-Based Features

- These features are those that aggregate information between specific pairs of hosts over a given period, taking into account the bi-directional flow of data:These features are those that aggregate information between specific pairs of hosts over a given period, taking into account the bi-directional flow of data:
- Flow Duration: Knowing how long it took to send the first and last packet of stream helps determine the types of connections that occur. Prolonged flow periods might prove either an accidental situation in which the bad link is active for a long time or a well-orchestrated attempt to evade monitoring and intel's notice for a long period.
- Total Flow Bytes and Packets: The volume, which is measured in both bytes and packet is through which malicious communications can be detected. Volumes may indicate that data is leaking out or they may signify that the resources are being thwarted with flooding attacks.
- Flow Rate: Flow rate, which is a measurement of the number of bytes or packets per second and can be used to spot traffic surges, is characteristic of DDoS (Distributed Denial of Service) attacks, where the service is disrupted resulting from a targeted system being flooded with inbound requests to slower and disrupt performance.

Characteristically, flow-based flows have proven to be the dominant factor in the relationship between traffic and networks and this is shown in the successful identification of anomalous flows that hint at malware or cyber-attacks. Our way of performing detailed and advanced encoded statistics and flows here show the detain of these behavior and their anomalies. Thus, a holistic approach is put in palce for incorporating precision into a machine learning models that is further used for detection of malware and that serves as a reliable barrier against threats of cyber security.

5. Feature Scaling

Feature scaling leads to rescaling feature values which may cause inappropriate comparison if it has not been normalize. Therefore, feature scaling helps to standardize the feature values and improves the accuracy of the machine learning model by enabling the model to compare the features better.

6. Designing the final feature set from the given list.

Following the selection of feature sets, a list that involves encodings of categorical data was created in addition to statistical and flow-based features for machine learning-based analysis was compiled.

7. Machine Learning Based Malware Detection

We employed three outstanding machine learning algorithms to detect and classify malware embedded with network traffic data.These methods incorporate various aspects of data and techniques of machine learning-based learning-based supervised learning, unsupervised learning, and timeseries analysis. The ensuing narrative elucidates each method and delineates its implementation in our inquiry:The ensuing narrative elucidates each method and delineates its implementation in our inquiry:

Anomaly Detection Using Clustering The method uses the clustering technology in order to perform the analysis of the network traffic data, to reveal the natural patterns and abnormalities which indicate malware activities. Characters like 'Packet length', 'Source port', 'Destination port', 'Protocol' and 'Traffic type' are used in the process out of which 'Packet length', 'Source port', 'Destination port' and 'Traffic type' filter a bulk of the The way in which malware-related anomalies exist is through either deviating from the main traffic patterns, or exhibiting concentrations that the network anticipates. The name of the game is virus hunting by recognizing the novel or unknown malware strains as well as taking any preventive measures against the new ones still in the infancy.

Application in Research: Aiming at the focus points such as the K-Means algorithm, our analysis enabled to pull out irregular clusters. Ultimately the resultant clusters underwent a careful scrutiny, which was

narrowed down to the clusters that are smaller in size or found in isolation, which indicates that they may also harbor innovative malware distribution methods.

Classification Using Supervised Learning:

The application of this methodology (which is among the few others present in this section) is noticeable mostly when the data set comprises labels showing the presence of malware. The way to obtain the knowledge is by using markers like 'IDS/IPS Alerts'. Utilizing supervised learning models allows the unknown nature of features such as 'Protocol', 'Packet Length', 'Action Taken' and 'Severity Level' to be learnt by the model and convey what would be a malware signatures pattern within the data. Follows, which is to say the model has learned to classify new cases on the basis of this acquired pattern recognition ability, thus making the model work better in the identification of the known types of malware.

Application in Research: To build our classification and regression load we used Support Vector Machines (SVM) and Neural Networks applied to our annotated dataset. The performance of the suggested models is evaluated with the help of some selected metrics such as accuracy, precision, recall, and F1 scores to detect how well the models can identify malware.

Time Series Analysis for Malware Detection

Acknowledging the temporal dynamics often associated with malware activity, we engaged in time series analysis. This approach scrutinizes time-stamped features like 'Timestamp', 'Packet Length', and 'Packet Type' across intervals to unearth patterns or surges in activity symptomatic of malware incursions, such as DDoS exploits or scanning operations. Long Short-Term Memory (LSTM) networks have an inherent capacity to encapsulate long-term dependencies in time series data, and thus stand out as particularly suitable for this analysis.

Application in Research: Our study harnessed LSTM models to thoroughly study network events, with the aim of unveiling distinctive patterns indicative of malware operations. This technique enabled the identification of potential malware occurrences by vigilant monitoring for deviations from established traffic paradigms.

Collectively, these machine learning approaches give our research a multifaceted lens through which malware detection is not only feasible but also markedly precise. Integrating these diverse methodologies, our research aims to provide comprehensive network security, substantially enhancing the detection capabilities against an expansive spectrum of malware entities with elevated accuracy and operational efficiency.

3.1 Anomaly Detection Using Clustering

The first ML technique applied in the study is unsupervised clustering which plays a role of the K-means algorithm in the malware analysis to gain patterns from cybersecurity data attacks. This particular technique entails selecting the features that are relevant for the dataset like length of the packets and protocol type, encoding non-numerical data, and normalizing features so that there is uniformity of scale across the data. The effectiveness of the clustering is evaluated using three metrics: for instance, the result of applying the Silhouette/Score, Davies-Bouldin Index, or Calinski-Harabasz/Index represents how close, or how far, the clusters are from each other while maintaining their internal cohesion. Following the analysis, the output is visually interpreted which lets us to identify an optimal clustering configuration. This would be critical to the identification of the most common attack patterns and anomalies, deepening the understanding of malware threats by offering vital information for detection and understanding.

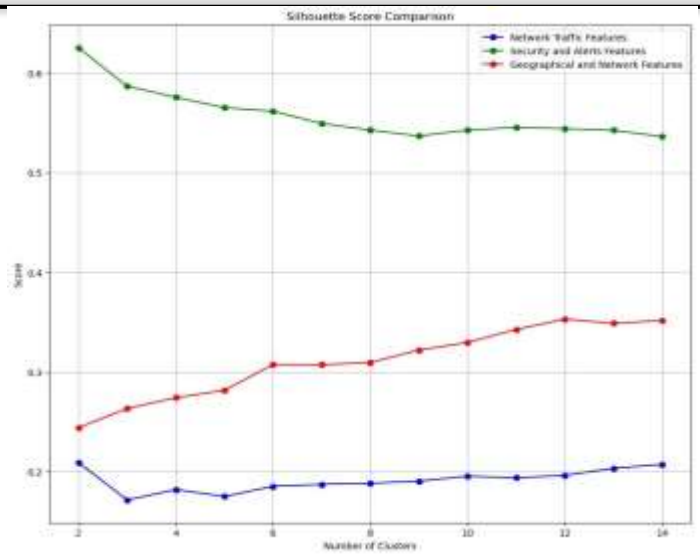

Figure 1 A graph showing Silhouette Score Comparison across different clustering levels for 3 classes of network traffic using 3 different feature sets

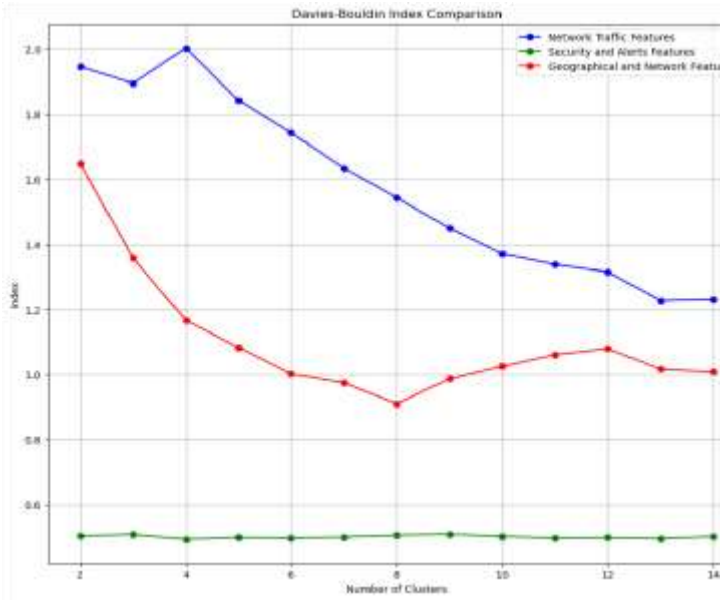

Figure 2 A graph showing Davies-Bouldin Index Comparison across different clustering levels for 3 classes of network traffic using 3 different feature sets

During the training phase, the Random Forest Classifier uses a solemn methodology in building decision trees based upon different features groups with the sole purpose of evaluating their influence on models performance. This stage is crucial for discovering network patterns and security errors, generally basic for reliably predicting the future cybersecurity threats or attacks. During the testing stage, the model is evaluated with fresh data which was not seen by training to get an idea of its performance in identifying new types of attack patterns.

Precision: This metric will help measure the accuracy of predicting positive outcomes across every attack¬† category by reducing false positives.

Recall: It has to do with how well the configuration may identify all instances of each class, which is a requirement for thorough threat detection

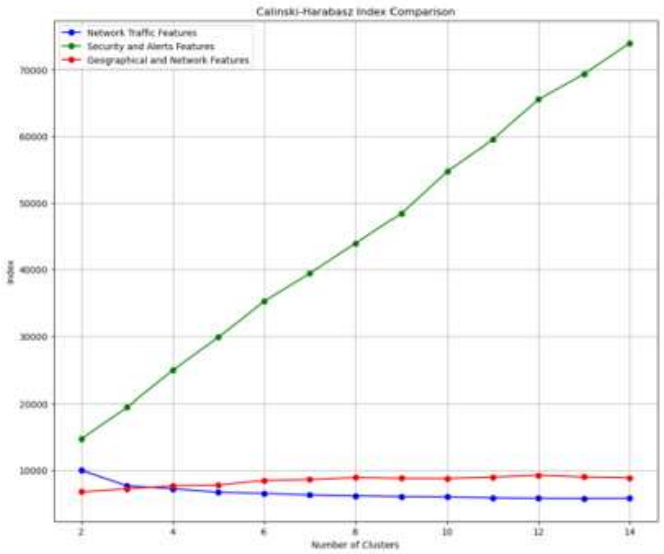

Figure 3 A graph showing Calinski Harabasz Index Comparison across different clustering levels for 3 classes of network traffic using 3 different feature sets

3.2 Classification Using Random Forest Classifier

With this approach another ML model used is the Random Forest classifier which is commonly used for making highly accurate and consistent predictions in cybersecurity attack data analysis. This algorithm could be observed to increase the accuracy of predictions by selecting the class which may be observed most among multiple decision tree outputs, a method that is particularly effective in cyber-security which may be complex and variable.

F1-Score: Acting as a joint factor between Precision and recall, F1-score is especially useful in cases that the class distribution goes skewed and gives more specific assessment of the classifier. Support: This measure, however, is indicators of actual attack occurrence frequency for each class and inject into the model performance across various attack types. To further improve the model's performance, the ensemble technique inherent to the Random Forest plays a vital role in preventing overfitting. This is achieved by averaging or taking the mode of predictions across multiple decision trees, thereby enhancing the reliability of the predictions. Moreover, the standardization of features and encoding of categorical variables are emphasized as essential steps to increase the model's interpretability

of data, a necessity for applying machine learning in cybersecurity effectively.

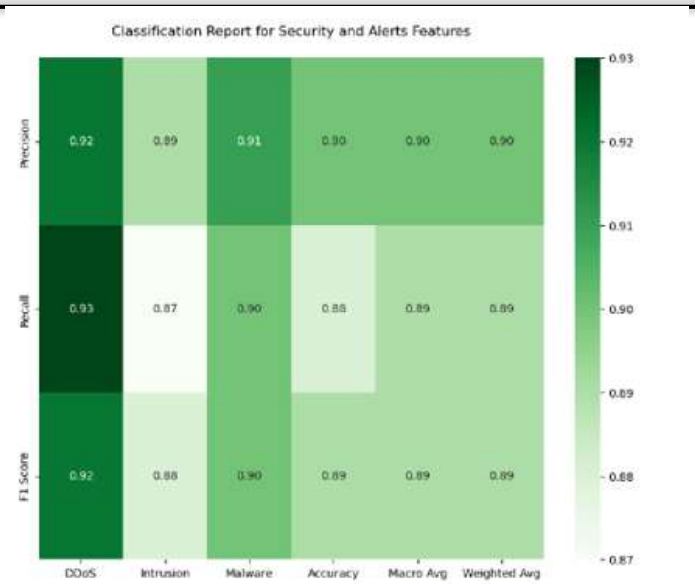

Figure 4 A heatmap showing the distribution of 3 evaluation metrics for the security and alerts features

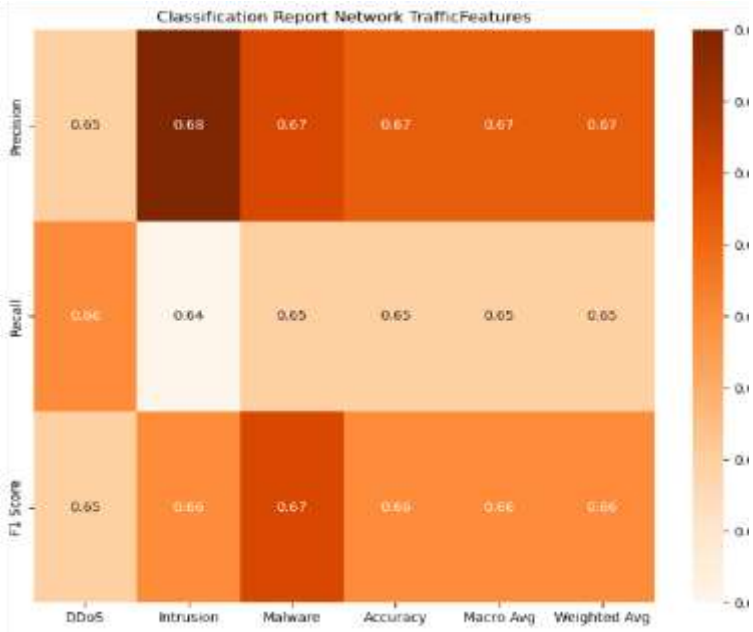

Figure 5 A heatmap showing the distribution of 3 evaluation metrics for the Network Traffic features

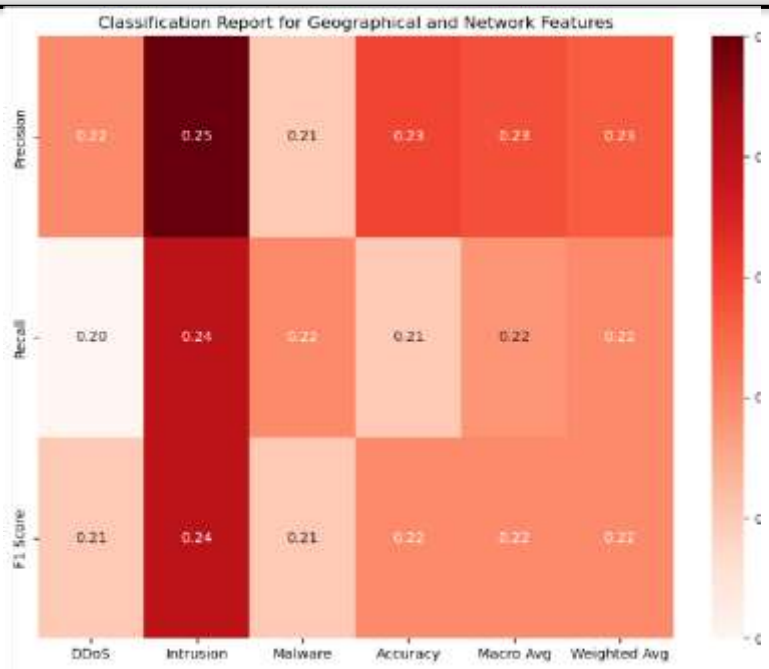

Fig. 6. A heatmap showing the distribution of 3 evaluation metrics for the Geographicla and Network features

3.3 LSTM Based Time Series Analysis For Malware Detection

The third approach is to use the Long Short-term Memory (LSTM) network which is good at sequential data. The model's complexity is optimized by using hyperparameters as the number of units in the LSTM layers and the number of epochs. Altering these scalar parameters determines how well the model learns.

Model's "validation accuracy" is also measured, indicating the ratio of true predictions it makes for the new data. It provides a basis for determining whether this technique might work in real-world situations.

Hyperparameter Tuning:

Adjusting hyperparameters refers to training diverse LSTM models with varied hyperparameters before finally selecting the best setup which optimizes data patterns without overfitting to the training data.

The comparison of various configurations helps us to find the best solution in which fit the data patterns effectively and prevent overfitting. It is a good compromise between high accuracy and robustness in the presence of new data, and therefore, it is a method suitable in cybersecurity for predicting network attacks.

The consequences of several configurations are plotted on a graph where the number of LSTM units, the number of epochs, and the validation accuracy can be seen. A picture help to grasp patterns of the data (underfitting) and too much specialization on the training data (overfitting).

Model Selection The graph is able to illustrate the specific model configuration that provides the best accuracy without excessive overfitting. The proposed strategy achieves a balance between efficient and sufficient learning, by which a cybersecurity system is developed and employed in predicting network attack types. Through this iterative process of this training and evaluation, we are increasing the reliability and predictive quality of the model making it an indispensable cybersecurity tool for the professionals in the ongoing fight against network threats.

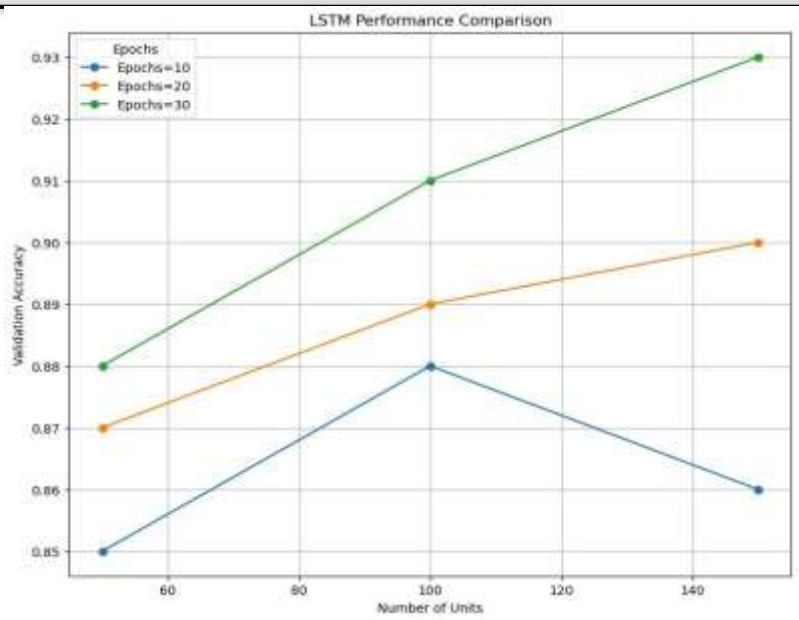

Fig. 7. The line graph shows validation accuracy at unit levels for different epochs of the algorithmic run

4. Results

Our research focused on the machine learning-oriented cybersecurity enhancement notably around the detection of DDoS, Intrusion, and Malware intrusion vectors. Three distinct analyses were conducted: an evaluation of clustering model indices, performance assessment of Random Forest classifiers through heatmaps, and optimization of LSTM networks via epoch numbers and unit counts.

The Geographical and Network Features set was revealed to be the best group of variables by the K-means model evaluation. The variables produced clusters that had high quality as indicated by the high Calinski-Harabasz Index and Silhouette Scores. It indicates that efficient analysis of the similar kinds of cybersecurity threat data can be made. On the other hand, the Secure and Alerts Features exhibit the best spread of the clusters implying its utilization in distinguishing different kinds of cyberattacks, since it is indicated by a low Davies-Bouldin Index. The Network Traffic Features had its performance evaluated in terms of the two aspects – clustering and attack patterns – and lagged with outcomes left to be desired on these measures. High in classification heatmaps, Component Random Forest classifiers also distinguished the Security and Alerts Features set as a high precision, recall, and F1-scores set. This is an evidence that effectiveness in recognizing a wide range of cyberattacks is much higher than the rest. Medium improvement was performed by the Network Traffic Feature set in compare to the Network and Geographical Features set which did not obtain good results that means it not effective for more accurate classification of attacks. The LSTM network optimization showed, that adjusting the training epochs and unit numbers is required to be done. Over the epochs, the validation pattern showed that an increase in the number of epochs usually improved the accuracy, which clearly indicates that longer training is beneficial. The best condition available was 100 LSTM units in 30 epochs which further improved the model without severe overfitting. But adding 100 units at 30 epochs signaled at overfitting phenomenon that make the model gains complexity and loses generalizability.

5. Conclusion

This study evaluated various machine learning techniques for malware detection, highlighting the importance of effective feature selection. Among the models tested, Security and Alerts Features consistently delivered the highest precision and recall. Random Forest classifiers and well-tuned LSTM networks showed strong performance, provided they were properly optimized to avoid overfitting. Overall, feature extraction and model tuning are critical to

enhancing the accuracy and adaptability of ML-based cybersecurity systems.

REFERENCES

Vinayakumar, R., Soman, K. P., and Poornachandran, P., "ScaledMalNet: A scalable deep learning framework for malware classification," Journal of Cybersecurity and Information Management, vol. 5, no. 3, 2019, pp. 120–132.

Hajraoui, M., and El Merabet, Y., "Feature selection and classification techniques for malware detection using machine learning," Procedia Computer Science, vol. 150, 2019, pp. 586–591.

Chintha, R., Basha, K., and Alomari, A., "A comparative analysis of machine learning algorithms for malware detection," in Proc. 2020 Int. Conf. on Computer, Communication and Signal Processing (ICCCSP), Chennai, India, 2020, pp. 108–113.

Moubrack, M., and Feghali, K., "Performance analysis of random forest classifier in malware detection," International Journal of Network Security & Its Applications (IJNSA), vol. 12, no. 4, 2020, pp. 45–54.

Machuche, F., Abouchabaka, L., and Bellafkih, M., "Malware detection using traditional machine learning algorithms and hybrid techniques: A survey," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 11, no. 9, 2020, pp. 134–140.

Sheneamer, A., Alharbe, N. A., and Alqarni, A. A., "Visual malware detection using image processing and machine learning," IEEE Access, vol. 10, 2022, pp. 2853–2865.

Alqahtani, A., "Application of machine learning models in cybersecurity: A focus on malware detection," International Journal of Computer Applications, vol. 183, no. 18, 2021, pp. 30–35.

Rkhouya, H., and Chougdali, M., "Random Forest based malware detection in large datasets," Journal of Information Security Research, vol. 9, no. 2, 2021, pp. 97–105.

Odintsov, O., Ivanov, A., and Petrov, S., "Detection of polymorphic malware using DT, CNN, and SVM classifiers," in Proc. 2022 Int. Conf. on Cybersecurity Trends (ICCT), Moscow, Russia, 2022, pp. 55–61.