A Robust Ensemble Learning Approach for Effective Malicious Traffic Detection in Mutli-Environment Networks

Umar Farooq Khan
MSc Artificial Intelligence
National College of Ireland
ID: x22179780

***Abstract*—The emergence of AI-based cyber attacks poses a significant challenge to existing detection systems, necessitating the adaptation of newer systems to detect these new threats. Additionally, the proliferation of newer types of devices and architectures further complicates the cybersecurity landscape. At present, there exists a gap in the capability of detection systems to effectively address both AI-generated attacks and the diversity of network architectures. It is fairly impossible to segment the network, hence to meet these challenges, we require one versatile detection system for identifying malicious traffic across diverse network types, newer architectures, AI-generated threats. However, existing datasets lack representation of multi-environment networks. To address this, we utilize the M-En dataset which represents both; traditional IP-based and IoT traffic. Our methodology involves applying Bi-Directional GRU and Bi-Directional LSTM and then stacking them on top of each other (BiGRU-BiLSTM), then further ensembling them to improve the robustness of the model which indicated that the best performing generated model achieved an accuracy of 0.972, a precision of 0.986 in another ensembled model while BiLSTM gave 0.98 accuracy. All the models have advantage on each other. These findings underscore the efficacy of ensemble learning approaches in enhancing the detection capabilities of multi-environment traffic detection systems.**

***Index Terms*—Ensemble Learning, Cyber security, Cyber intrusion detection system Network security, Cyber attack detection system, AI Generated cyber attacks, Malicious Traffic Detection**

I. Introduction

In today's rapidly evolving digital landscape, the threat of cybercrime looms larger than ever before. With cybercriminals perpetually innovating and deploying increasingly sophisticated attacks, organizations worldwide find themselves in a perpetual game of catch-up to defend against these threats. Information leakage is the growing and concerning problem for private or government organizations, or private companies [13] which means the release of confidential data leads to significant financial losses and damages to the reputation of companies [14]. Both Supervised and Unsupervised Machine learning can be used for making these system [15] but we will be using Supervised Machine Learning and treating it a binary classification problem. Traditional cybersecurity solutions, such as Network Intrusion Detection and Prevention Systems (IDS/IPS), while still in use, are facing diminishing returns in the face of complex network infrastructures and applications. This necessitates a shift towards more adaptive and proactive defense mechanisms. Consider Machine Learning (ML) as a powerful tool that could help us fight against cyber threats. ML algorithms, empowered by accessible hardware and computational resources, possess the capability to sift through massive datasets and discern subtle patterns indicative of malicious activities. Through supervised learning techniques like classification, ML can effectively differentiate between benign and malicious network traffic, thus making our model best for our defenses against cyber attacks. Moreover, ML's inherent flexibility positions it as a promising solution for tackling the ever-evolving landscape of cyber threats, including those stemming from the proliferation of Internet of Things (IoT) devices and traditional IP-based networks. As the frequency and complexity of cyber attacks continue to surge, it is imperative that organizations invest in robust detection systems that harness the power of AI and ML. These systems not only enable organizations to mitigate the risks posed by cyber threats but also empower them to anticipate and preemptively counter emerging threats, ensuring a safer digital environment for all stakeholders.

The report is divided into three parts: Introduction, where we discuss the problem, what we're trying to find out, and what the report covers; Literature Review, where we look at what other people have done in this area; and Methodology and Evaluation, where we explain how we're going to do our study, and we evaluate the results.

II. Related Work

With the advancements in technology, newer devices and its new architectures are introduced in the market every month. It means newer hardware along with the software, networks, protocols, architecture are also introduced and improved very frequently. One exmaple is the IoT devices which became very popular since the introduction of 5g networks. Now a days our home, office, cars everything is connected with these devices and this exponentially growth and reliane can directly affect users and citizen. [16]

A. Detecting Malicious Traffic for Traditional IP-Based Traffic

In [9] the researchers explored various Machine Learning (ML) models, aiming to identify the most effective approach

for detecting malicious traffic. Among the models tested on the UNSW-NB15 dataset, Random Forest (RF) emerged as the top performer, achieving an impressive accuracy of 98.96%. Alongside RF, the study also evaluated models such as XGBoost, K Nearest Neighbour (KNN), and Deep Multi-Layer Perceptron (Deep MLP), the latter being a deep learning model with sequentially connected layers containing multiple neurons. Despite the complexity of the dataset, the accuracy attained by these models was notably high, indicating their potential utility in detecting specific types of malicious traffic. These findings underscore the importance of employing robust ML techniques in bolstering network security and mitigating potential threats. [10] proposed the utilization of BiDirectional Gated Recurrent Unit (Bi-GRU) models for analyzing the CICIDS2017 and UNSW-NB15 datasets. Impressively, their implementation yielded high accuracy rates of 96.84% and 95.68% for the respective datasets. This innovative approach showcases the effectiveness of Bi-GRU in accurately processing and interpreting complex network data, underscoring its potential as a robust tool for intrusion detection and network security analysis.

In the study referenced as [11], researchers employed a range of machine learning (ML) algorithms, such as AdaBoost, GRU, LSTM, MLP, and DT, for binary classification tasks. However, they encountered challenges due to dataset imbalance. To address this issue, they conducted a feature selection process using weighted Random Forest (RF) and Gini Impurity methods. Remarkably, applying these techniques led to notable improvements in performance metrics. This underscores the significance of employing appropriate feature selection methodologies to enhance model effectiveness, particularly when dealing with imbalanced datasets. These findings emphasize the importance of robust preprocessing techniques in optimizing ML algorithms for binary classification tasks. In their study [12], researchers proposed a novel approach to Network Intrusion Detection Systems (NIDS) utilizing deep learning methodologies. They constructed a NIDS employing a sparse auto-encoder and softmax regression, with a focus on anomaly detection. Evaluation was conducted using the NSL-KDD benchmark dataset to gauge detection accuracy. The findings revealed superior performance compared to existing NIDSs, particularly in distinguishing between normal and anomalous network behavior on test data. To further enhance performance, the researchers suggested incorporating techniques like Stacked Auto-Encoder for unsupervised feature learning, alongside classifiers such as NB-Tree, Random Tree, or J48 for classification tasks. Additionally, they outlined plans to develop a real-time NIDS for practical network deployment using deep learning techniques and explore methods for on-the-fly feature learning directly from raw network traffic headers in future investigations.

Several research papers [10] implemented the BiDirectional Gated Recurrent unit (Bi-GRU) on the CICIDS2017 and UNSW-NB15 datasets, attaining accuracies of 0.968 and 0.9568 correspondingly.

B. Detecting Malicious Traffic for IoT-based Traffic

[4] has conducted a comparative analysis between deep learning (DL) and traditional machine learning (ML) algorithms utilizing the CICIDS2017 dataset. Results highlight the superior performance of DL models, particularly LSTM and CNN+LSTM architectures, achieving accuracy rates ranging from 96.24% to 97.16% in cyber threat detection. In contrast, SVM, Bayesian classifiers, and random forests demonstrate lower accuracy rates ranging from 94.64% to 95.5%. Although the 1D-CNN model demonstrates comparable accuracy at 95.14%, the MLP model trails significantly behind at 86.34%. There is a need to explore interpretability and scalability issues in large-scale environments, along with the integration of ensemble learning techniques to further enhance intrusion detection system (IDS) performance. [5] discuss the application of generative deep learning methodologies, particularly Adversarial Autoencoders (AAE) and Bidirectional Generative Adversarial Networks (BiGAN), in cyberattack detection utilizing the IoT-23 dataset.They are using Random Forest and KNN as baseline models, the research compare their efficency against AAE and BiGAN models. Strikingly, both AAE + KNN and BiGAN + KNN configurations demonstrated good accuracies and F1-scores. Notably, both AAE and BiGAN models achieved exceptional F1-Scores of 0.99, with BiGAN further trained to identify unknown attacks, including novel zero-day threats, exhibiting an F1-Score ranging from 0.85 to 1.

[6] present a novel wireless IDS approach, integrating FFDNN with WFEU, evaluated on AWID and UNSW-NB15 datasets. WFEU, utilizing Extra Trees, optimizes feature vectors, enhancing detection accuracy and computational efficiency. Comparative analysis against RF, SVM, NB, DT, and kNN reveals remarkable accuracies, with WFEU-FFDNN achieving 99.66% and 99.77% overall accuracies for binary and multiclass tasks. Acknowledging limitations in class detection rates and WFEU impact, further research is vital for optimization and validation across diverse datasets. [7] discuss the advancements in IDS for 5G and IoT networks highlight deep autoencoded dense neural network algorithms as effective tools against evolving security threats. Using the diverse AWID dataset for training and evaluation, the study achieves an impressive 99.8% overall detection rate for various attacks. Despite success, challenges such as false positives and false negatives persist, urging future refinement efforts. Additionally, a data mining-based hybrid IDS shows promise in mitigating threats but faces similar accuracy challenges. These findings emphasize the ongoing necessity for optimized IDS to safeguard 5G and IoT networks.

This study [8] delve into the details of IoT attackable packet sequences, focusing on both internal and external packet origins. Their proposed defense strategy involves leveraging IoT-enabled network services and deploying a custom artificial neural network (aNN) architecture designed specifically for Raspberry Pi. Through precise simulations across multi-threaded Raspberry Pi networks, the study aims to strengthen

the detection and prevention capabilities against a range of IoT attacks. The study underscores significant improvements in attack detection and prevention facilitated by the novel aNN structure. This research contributes a tailored aNN framework optimized for IoT environments for smart IoT devices.

C. Gap and Summary of Previous Work

Recent research has delved into detecting malicious traffic within various IoT and traditional IP-based datasets, employing ensemble models and meticulous feature engineering to achieve notable accuracy. However, a persistent challenge has emerged: the absence of a unified model capable of effectively addressing both IP-based and IoT network architectures. Prior studies often overlooked the issue of imbalanced datasets, which plagued many of the datasets utilized.

Researchers have extensively explored traditional IP-based and IoT datasets, harnessing machine learning algorithms, notably favoring ensemble learning approaches, to discern abnormal and malicious traffic with remarkable precision. Yet, the inherent complexity of Multi-Environment architectures, encompassing both Tr.IP and IoT networks within a singular network infrastructure, remains largely unaddressed. Segregating the network or relying on VLAN segmentation, which is costly also presents security concerns. Furthermore, the oversight of imbalanced datasets in previous studies underscores the necessity for a novel approach, leading to the adoption of the M-En dataset in our research endeavors.

III. Methodology

A. Dataset Description

In order to make a AI based secure system that detect malicious traffic in M-En, we should use that dataset which have characteristics of both types of networks. Hence, a newer dataset M-En has been made. Our study used M-En networks which is made from the combination of Traditional IP-based traffic (UNSW-NB15 [26]) and IoT-based traffic (IoTID-20 [29]) through various techniques. M-En is only 25% of the combined datasets and has 30,000 records, having 20 categorical features and 1 target feature showing if the record is a potential Malicious or Normal.

TABLE I
Class Counts

Target	Count
Anomaly	22459
Normal	7541

Figure 1 showing top 10 Most correlated features and 15-14 and 17-16 are the most correlated.
Cybersecurity, intrusion detection, and malicious traffic detection systems play pivotal roles in safeguarding computer networks and systems from unauthorized access, data breaches, and malicious activities. Figure 2 shows the generic architecture of our system where traffic from different types of network types (M-En) are coming and is being checked by

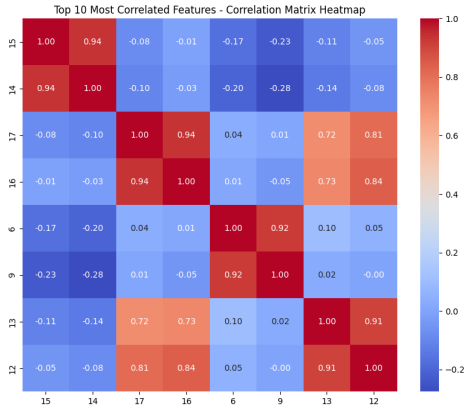

Fig. 1. Top 10 Most Correlated Features

our system by our proposed system and it will classify and alert the system further, if the traffic is suspicious or normal which usually Intrusion detection system or Malicious Traffic Detection Systems does.

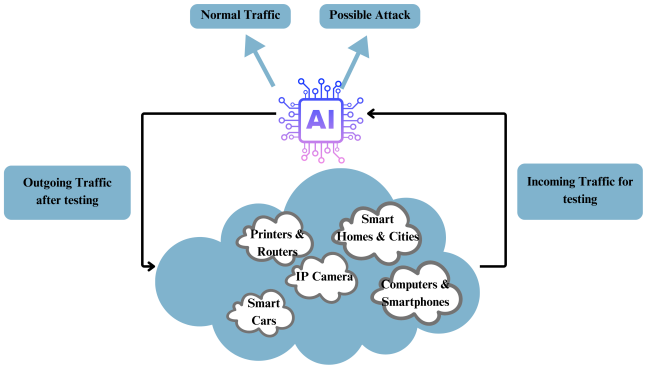

Fig. 2. Proposed Detection System Architecuture

This section discusses about the proposed methodology which is shown in figure 3 below. We used such deep learning models which has the potential of producing good results on our dataset; used Bidirectional Gated Reccurent Unit Bi-GRU), Bidirectional Long Short Term Memory (LSTM) and its combination that is Bi-GRU and Bi-LSTM and then apply ensemble approach and evaluate results on finding out which performs better for which type of metric. BiGRU and BiLSTM models excel in classifying cyber attacks on continuous data streams due to their capacity to capture long-term dependencies, leverage bidirectional processing, and integrate gating mechanisms. Hence, these models and its combination will be used. These architectures autonomously learn pertinent features from raw input data, diminishing the need for manual feature engineering and enhancing adaptability to evolving attack patterns because each and every feature is important in this domain. Their flexibility and robustness make them potent tools for precisely identifying and classifying various cyber attacks, making them good for our use case which relies on continuous data.

Our methodology starts with doing preprocessing on the M-En dataset which is scaling, we used Standard Scaling, then we apply GRU, BiLSTM, BIGRU separately and them ensemble them to check if they perform better or not. Then based on the results we evaluate all the models including ensembled model with metrics like accuracy, precision, recall, f1 score and ROC Curve which further adds to our understanding which models perform best and whether ensembling is a good approach for our use case. Figure X shows our methodology.

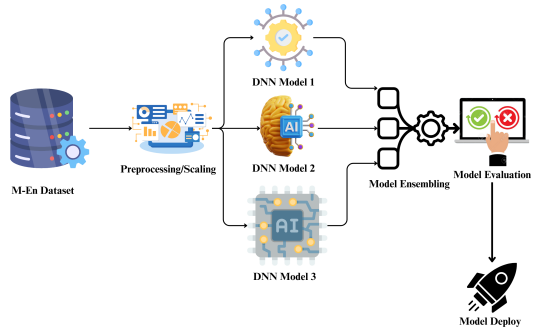

Fig. 3. Proposed Methodology

BiGRU is a significant improvement in recurrent neural networks (RNNs) that tackles the issue of losing important context information in traditional models. It is a variant that of LSTM and GRU. BiGRU uses bidirectional processing and gated mechanisms to better capture long-term patterns in sequential data as shown in 4. By running two GRUs

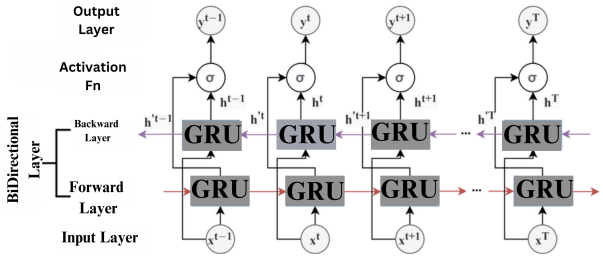

Fig. 4. Bi-GRU Architecture

in opposite directions, BiGRU can process input from both past and future contexts at the same time, improving its grasp of temporal relationships. It combines information from both directions to create a more complete picture, enhancing its ability to predict outcomes accurately across a range of sequential tasks. Overall, BiGRU represents a step forward in making RNNs more effective and understandable for various applications. This bidirectional processing paradigm not only enhances the model's grasp of temporal relationships but also ensures a nuanced capture of context information crucial for precise predictions in various applications, marking a significant advancement in the field of RNNs.

Bi-LSTMs, an advancement over standard RNNs, adeptly address the challenge of capturing temporal dependencies by incorporating information from both past and future time steps. By splitting hidden neurons into forward and backward states, Bi-LSTMs extract comprehensive temporal information, eliminating the need for additional time delays characteristic of standard RNNs. This bidirectional processing capability enables Bi-LSTMs to consider contextual information from both preceding and subsequent timestamps simultaneously, enhancing their capacity to capture intricate dependencies and patterns within sequential data.

Operating as an extension of the Long Short-Term Memory (LSTM) architecture, BiLSTM further enhances its capabilities by incorporating bidirectional processing alongside its memory cell units, input gates, forget gates, and output gates. This dual-directional strategy involves running two distinct LSTM networks concurrently: one processing input sequences from past to future, while the other processes them from future to past. By seamlessly integrating bidirectional processing with LSTM's memory mechanisms, BiLSTM gains a holistic perspective on the input sequence, enabling it to make more informed predictions across a wide array of tasks.

Following the bidirectional processing of the input sequence, BiLSTM combines the outputs of the two LSTM networks, leveraging insights extracted from both past and future contexts. Combining the outputs helps BiLSTM better understand the input, making its analyses and predictions more detailed. Consequently, BiLSTM excels in tasks requiring context-aware predictions, such as sequence labeling, sentiment analysis, and speech recognition. In summary, BiLSTM represents a sophisticated advancement in RNN architectures, offering enhanced capabilities in capturing long-range dependencies and making informed predictions based on a holistic interpretation of sequential data.

In the proposed detection system, ensemble learning is utilized to merge the predictions of Bi-GRU and Bi-LSTM models to enhance overall performance. This ensemble approach facilitates the capture of a wider spectrum of features and patterns present in network traffic data, thus enhances the system's ability to identify malicious activities. Furthermore, ensemble learning contributes to model robustness and generalization by mitigating overfitting risks and incorporating diverse perspectives on the data. The resulting ensemble model yields more stable and reliable predictions compared to individual models, underscoring its efficacy in mitigating the impact of model variance and errors. Ensemble learning boosts accuracy by combining predictions from diverse models, leveraging their varied perspectives and mitigating individual model biases. It diminishes overfitting risks by averaging predictions and corrects errors through consensus, resulting in more robust and reliable predictions. This collaborative approach enhances accuracy by harnessing the collective intelligence of multiple models, surpassing the performance of individual models in isolation.

IV. Evaluation

Once the models are properly implemented, they are evaluated based on accuracy, recall, precision, F1 score, G-Mean,

True Negative Rate, loss, AUC-ROC value, ROC Curve, number of correct predictions, false negatives, true positives, and the Jaccard Index. The Jaccard Index is calculated as the ratio of the number of correctly predicted positive labels to the total number of unique labels across both the predicted and true sets:

$$JaccardIndex = \frac{TP}{TP + FP + FN}$$

Where G-Mean which is useful when classes are imbalanced and is a square root of product of sensitivity and specificity.

A. BIGRU

The evaluation results reveal impressive performance metrics for Bi-GRU. It achieved an accuracy of 0.986 and an AUC score of 0.993, indicating excellent discrimination ability between positive and negative classes. Figure 5 shows a spike at the starting but gradually dropping as the epochs are increasing.

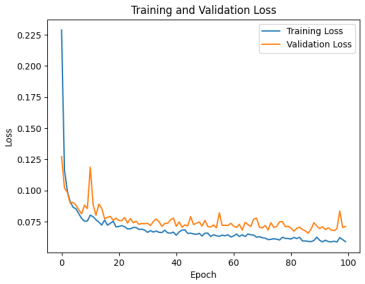

Fig. 5. Loss Curve for BiGRU

The confusion matrix shows a low number of false positives and false negatives. Furthermore, the classification report demonstrates high precision 0.96, recall 0.94, and F1-score 0.95 for the positive class, and similarly high values for the negative class, indicating robust performance across both categories.

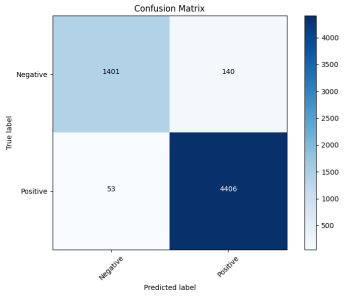

Fig. 6. Confusion Matrix BiGRU

The ROC curve illustrates in Figure 7 strong performance across various thresholds, with a true positive rate of 0.95 and a false positive rate of 0.025.

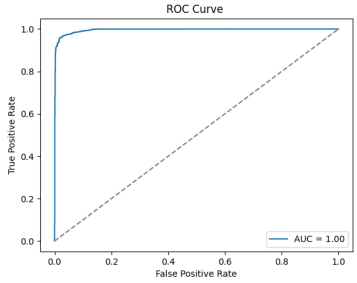

Fig. 7. ROC BiGRU

B. BILSTM

Our Bi-LSTM comprises a bidirectional LSTM layer with 256 units, followed by a dense layer with sigmoid activation. This architecture totals 133,377 trainable parameters. Over 100 epochs, the model consistently improves its performance. Notably, the training accuracy reaches approximately 0.972, while the training loss decreases to around 0.057. Meanwhile, the validation accuracy achieves roughly 0.971, with a validation loss of about 0.0626 as shown in Figure 8

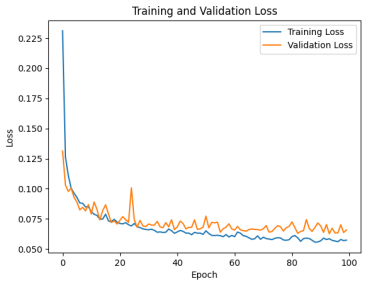

Fig. 8. Loss Curve for BiLSTM

This suggests that the model generalizes well to unseen data, as evidenced by the comparable performance on the validation set. The classification report shows that the precision is 0.94 for class 0 and 0.98 for class 1, the model demonstrates high accuracy in correctly predicting instances for both classes. Similarly, recall scores of 0.95 for class 0 and 0.98 for class 1 indicate the model's ability to identify a large proportion of true instances for each class. The high F1-scores of 0.94 and 0.98 for classes 0 and 1, respectively, signify a balanced trade-off between precision and recall. Furthermore, the macro and weighted averages of precision, recall, and F1-score all hover around 0.96 to 0.97, indicating consistent performance across classes and effectively handling class imbalance. With an accuracy of 0.97, this model correctly predicts outcomes in nearly all cases. Its precision, measuring the accuracy of positive predictions, is notably high at approximately 0.981. Moreover, the recall, which gauges the model's ability to identify all relevant instances, stands impressively at around 0.978. The F1 score, a balance between precision and recall, is also robust, nearing 0.979. The confusion matrix reveals that

the model makes relatively few mis classifications, with 84 false positives and 96 false negatives as shown in Figure **??**

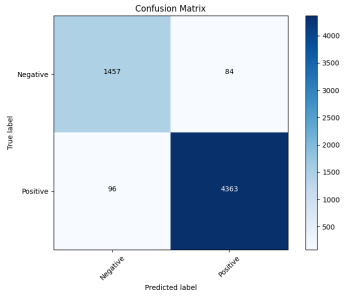

Fig. 9. Confusion Matrix BiLSTM

and showing an ideal ROC-Curve in Figure 10

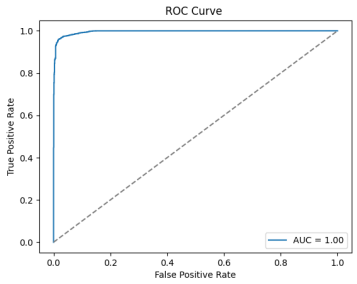

Fig. 10. ROC BiLSTM

C. LSTM

LSTM had several dense layers with rectified linear unit (ReLU) activation functions, followed by dropout layers to prevent overfitting. The output layer used a sigmoid activation function for binary classification. During training, the model underwent 100 epochs with a batch size of 32. To mitigate overfitting, dropout layers were incorporated after certain dense layers, with a dropout rate of 0.8. The Adam optimizer was employed, along with the binary crossentropy loss function. Across epochs, both training and validation accuracies progressively increased, indicating effective learning. The training accuracy reached approximately 96.22%, while the validation accuracy peaked at around 96.78%.

Similarly, as shown in Figure 11. the loss decreased steadily for both training and validation sets, indicating that the model was effectively minimizing errors and learning the underlying patterns in the data. Notably, both training and validation results exhibited consistency, suggesting that the model was not overfitting to the training data. After 20 epochs, the model is showing fairly less loss. The achieved accuracy of 0.9652 suggests strong performance. Additionally, the observed loss of 0.0681 for the training data and 0.0754 for the validation data indicates that the model's predictions are close to the actual values, implying robustness in its generalization capability.

Moreover, Figure 12 shows perfect the ideal ROC Curve.

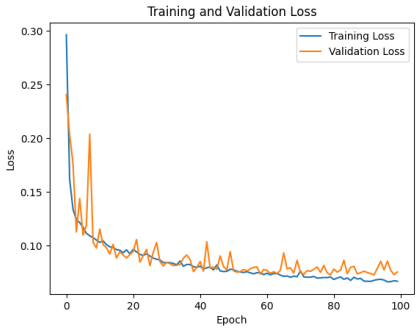

Fig. 11. Loss Curve for LSTM

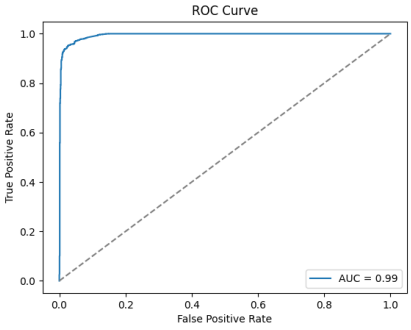

Fig. 12. ROC Curve LSTM

D. BIGRU-BILSTM

The model demonstrates impressive performance across multiple metrics. With an accuracy of approximately 97%, it showcases a high level of correctness in predicting class labels. Figure 13 shows confusion Matrix. Precision stands strong at around 0.975, indicating the model's ability to accurately identify positive outcomes when it makes predictions. Moreover, its recall score of approximately 0.98 reflects its effectiveness in capturing positive instances from the dataset. The F1 score, which strikes a balance between precision and recall, hovers around 0.978, further underlining the model's robust performance.

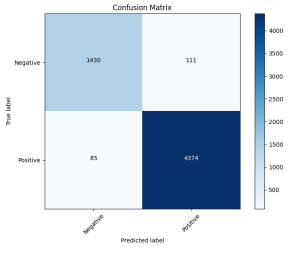

Fig. 13. Confusion Matrix BiGRU-BiLSTMTM

Figure 14 shows loss which slowly but gradually decreasing and reached 0.05 at the end of 100th epoch.

Figure 15 and AUC-ROC graph, a measure of the model's ability to distinguish between positive and negative instances,

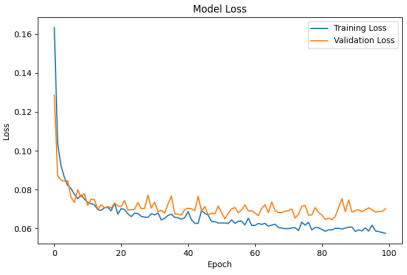

Fig. 14. Loss BiGRU-BiLSTM

achieves an outstanding score of 0.995, highlighting its discriminative power. Additionally, with a True Negative Rate (TNR) of about 0.927, the model excels in accurately identifying negative instances. The G-mean, balancing recall and TNR, reaches approximately 0.954, indicating a harmonious blend of sensitivity and specificity. Furthermore, the Jaccard Index, measuring the similarity between predicted and actual class labels, stands at about 0.957, reinforcing the model's overall efficacy in classification tasks. These collective metrics underscore the model's prowess and reliability in handling diverse datasets with high precision and accuracy.

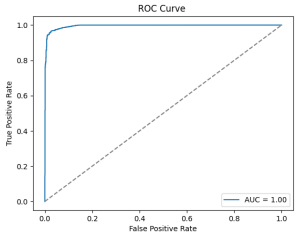

Fig. 15. ROC BiGRU-BiLSTM

Table II shows Classification report showing more better results for Anomaly class which is advantageous of our case.

TABLE II
CLASSIFICATION REPORT BIGRU-BILSTM

Class	Precision	Recall	F1-score	Support
0	0.94	0.93	0.94	1541
1	0.98	0.98	0.98	4459
Macro Avg	0.96	0.95	0.96	6000
Weighted Avg	0.97	0.97	0.97	6000

This table III shows all the metrics that we calculated in a tabular form for BiGRU-BiLSTM

E. Ensemble Models

The combination of 'BiLSTM' and 'BiGRU' stood out with the highest accuracy of 0.9717, closely trailed by the 'BiLSTM' single model at 0.9715, showcasing their efficacy in predictive tasks. Except for ('BiGRU', 'BiGRU-BiLSTM'), all combinations achieved accuracies surpassing 0.967, indicating strong overall performance. Precision-wise, ('BiLSTM', 'BiGRU') excelled with a 0.9860 score, emphasizing its ability

TABLE III
METRICS

Metric	Value
Accuracy	0.9673
Precision	0.9753
Recall	0.9809
F1 Score	0.9781
AUC-ROC	0.9953
True Negative Rate (TNR)	0.9280
G-mean	0.9541
Jaccard Index	0.9571

TABLE IV
ENSEMBLE MODEL METRICS USING HARD VOTING

Model Combination	Precision	Recall	F1 Score
BiLSTM	0.9790	0.9827	0.9809
BiGRU	0.9746	0.9812	0.9779
BiGRU-BiLSTM	0.9751	0.9843	0.9797
BiLSTM, BiGRU	0.9860	0.9758	0.9808
BiLSTM, BiGRU-BiLSTM	0.9842	0.9773	0.9808
BiGRU, BiGRU-BiLSTM	0.9804	0.9771	0.9788
BiLSTM, BiGRU, BiGRU-BiLSTM	0.9771	0.9841	0.9806

to minimize false positives. Similarly, 'BiLSTM' alone and ('BiLSTM', 'BiGRU-BiLSTM') displayed robust precision values above 0.978. On the recall front, ('BiGRU-BiLSTM') led with 0.9843, showcasing its capacity to minimize false negatives, while 'BiGRU' and ('BiGRU', 'BiGRU-BiLSTM') also showed strong recall above 0.977. For achieving a balance between precision and recall, ('BiLSTM', 'BiGRU') topped with an F1 score of 0.9808. Overall, the consistent performance across metrics underscores the efficacy of the 'BiLSTM' and 'BiGRU' ensemble, yet specific needs may dictate alternative combinations. Ensemble methods proved superior to individual models, underscoring the advantage of amalgamating diverse modeling approaches for enhanced predictive capabilities. Thus, the 'BiLSTM' and 'BiGRU' ensemble emerges as a formidable choice for various predictive tasks, albeit contingent on specific application requisites and objectives.

F. Experimental Setup

Experiments are done on Core i7 11th Generation machine with 16GB RAM and Integrated Intel Iris Xe Graphics using Jupyter Notebook and used libraries like Tensorflow, Keras, sci-kit Learn etc.

TABLE V
ENSEMBLE MODEL ACCURACY USING HARD VOTING

Model Combination	Accuracy
BiLSTM	0.9715
BiGRU	0.9670
BiGRU-BiLSTM	0.9697
BiLSTM, BiGRU	0.9717
BiLSTM, BiGRU-BiLSTM	0.9715
BiGRU, BiGRU-BiLSTM	0.9685
BiLSTM, BiGRU, BiGRU-BiLSTM	0.9710

V. Conclusions

In conclusion, our research presents a robust ensemble learning approach on top of sequential stacked BiGRU and BiLSTM for effective malicious traffic detection in multi-environment networks. Cybersecurity threats continue to escalate, posing significant risks to organizations and individuals worldwide. Traditional defense strategies are often outpaced by the complexity and frequency of cyber attacks, necessitating innovative solutions.

Our methodology leverages deep learning models such as BiGRU and BiLSTM, known for their ability to capture long-term dependencies and intricate patterns in sequential data. Through extensive experiments and evaluation, we demonstrated the effectiveness of our approach in detecting and mitigating malicious traffic in multi-environment networks.

The evaluation of the models based on various metrics demonstrates their performance across different architectures. Accuracy, recall, precision, F1 score, Jaccard Index, G-Mean, True Negative Rate, loss, AUC-ROC value, ROC Curve, number of correct predictions, false negatives, and true positives were utilized to assess the models' effectiveness. In all of the confusion matrices 'Positive' corresponds to 'Malicious' Traffic and 'Negative' corresponds to Normal traffic.

The evaluation is done on each and every model separately and evaluated all the possible combination of the models when we ensemble them which underscored the superior performances.

Bi-GRU gave accuracy of 0.986 and an AUC score of 0.993 signifying outstanding discrimination ability showing a true positive rate of 0.95, a classification report revealed high precision (0.96), recall (0.94), and F1-score (0.95) for both positive and negative classes, underscoring robust performance across both categories.

Bi-LSTM gave training accuracy of 0.972 and validation accuracy of 0.971. The model's accuracy of 0.97, precision of approximately 0.981, recall of about 0.978, and F1-score nearing 0.979.

When both Directional models are stacked on top of each other it gave an impressive accuracy of 0.98 for Malicious Traffic class, same in Recall, F1 Score. AUc-ROC Score is 0.995, with a G-Mean of 0.95, overall Recall is 0.98.

Overall, after ensembling in terms of accuracy (Bi-LSTM-BiGRU) achieved the highest accuracy that is 0.9860. (Bi-GRU-BiLSTM) achieved the highest recall of 0.9843. (BiLSTM-Bi-GRU) achieved the highest precision of 0.9860. Both the model combinations of (BiLSTM-BiGRU) and ensembled model of BiLSTM and Bi-GRU-BiLSTM achieved the highest F1 score of 0.9808.

Ensemble learning not only enhanced overall performance but also contributed to model robustness and generalization by mitigating overfitting risks and incorporating diverse perspectives on the data.

In conclusion, our research contributes to the ongoing efforts in cybersecurity by providing a novel approach to malicious traffic detection in multi-environment networks using Deep Learning and its Stacked models. By combining deep learning models and ensemble learning techniques, we aim to enhance the resilience of organizations against evolving cyber threats and safeguard critical digital assets.

A. Limitations Future Work

As a future work, we need to tackle the M-En dataset's limitations by finding ways to generate real M-En traffic. As cyber security domain is very sensitive hence original data must be employed. This would make the dataset more realistic, reflecting the complexities of hybrid networks better. Also, we should explore how well the S-DATE approach works outside of cybersecurity by testing it on different kinds of datasets on which M-En is based. In future, we can also include some other types of networks like SDN-based networks, to see if the approach works well across different setups. We could also improve the results by optimizing hyperparameters, adjusting network architecture layers, and incorporating advanced techniques like attention mechanisms or residual connections. Experimenting with larger and more diverse datasets, fine-tuning model parameters, or using transfer learning from pre-trained models could also improve performance. Additionally, we could also find some novel loss functions, regularization techniques may contribute to better model efficacy. Systematically evaluating these modifications and comparing performance metrics against baseline models will showcase the effectiveness of our architectural enhancements in boosting model performance. Also trying it out on lots of different and challenging datasets to see how well it can identify and deal with cyber threats.

VI. Dataset Availability

M-En dataset is publicly available and can be used for research purposes from this following link: https://www.kaggle.com/datasets/furqanrustam118/m-en-dataset.

References

[1] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[2] M. Roopak, G. Yun Tian, and J. Chambers, "Deep Learning Models for Cyber Security in IoT Networks," in 2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC), 2019, pp. 0452-0457, doi: 10.1109/CCWC.2019.8666588.

[3] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[4] Roopak, M., Tian, G.Y. and Chambers, J., 2019, January. Deep learning models for cyber security in IoT networks. In 2019 IEEE 9th annual computing and communication workshop and conference (CCWC) (pp. 0452-0457). IEEE.

[5] Abdalgawad, N., Sajun, A., Kaddoura, Y., Zualkernan, I.A. and Aloul, F., 2021. Generative deep learning to detect cyberattacks for the IoT-23 dataset. IEEE Access, 10, pp.6430-6441.

[6] Kasongo, S.M. and Sun, Y., 2020. A deep learning method with wrapper based feature extraction for wireless intrusion detection system. Computers & Security, 92, p.101752.

[7] Rezvy, S., Luo, Y., Petridis, M., Lasebae, A. and Zebin, T., 2019, March. An efficient deep learning model for intrusion classification and prediction in 5G and IoT networks. In 2019 53rd Annual Conference on information sciences and systems (CISS) (pp. 1-6). IEEE.

[8] Yoon, J., 2020. Deep-learning approach to attack handling of IoT devices using IoT-enabled network services. Internet of Things, 11, p.100241.

[9] Dhanya, K., Vajipayajula, S., Srinivasan, K., Tibrewal, A., Kumar, T. S., & Kumar, T. G. (2023). Detection of network attacks using machine learning and deep learning models. Procedia Computer Science, 218, 57–66. Presented at the International Conference on Machine Learning and Data Engineering

[10] Y. Feng and C. Wang, "Network anomaly early warning through generalized network temperature and deep learning," Journal of Network and Systems Management, vol. 31, no. 2, pp. 1–34, 2023.

[11] Disha, R.A. and Waheed, S., 2022. Performance analysis of machine learning models for intrusion detection system using Gini Impurity-based Weighted Random Forest (GIWRF) feature selection technique. Cybersecurity, 5(1), p.1.

[12] Niyaz, Q., Sun, W., Javaid, A.Y. and Alam, M., 2015, December. A deep learning approach for network intrusion detection system. In Proceedings of the 9th EAI International Conference on Bio-inspired Information and Communications Technologies (Formerly BIONETICS), BICT-15 (Vol. 15, No. 2015, pp. 21-26).

[13] Alneyadi, S., Sithirasenan, E., and Muthukkumarasamy, V., "A survey on data leakage prevention systems," *Journal of Network and Computer Applications*, vol. 62, pp. 137-152, 2016.

[14] Cheng, L., Liu, F., & Yao, D. (2017). Enterprise data breach: Causes, challenges, prevention, and future directions. Data Mining and Knowledge Discovery, 7(5), e1211.

[15] Delplace, A., Hermoso, S., & Anandita, K. (2020). Cyber attack detection using machine learning algorithms. arXiv preprint arXiv:2001.06309.

[16] Y. Li and Q. Liu, "A comprehensive review study of cyber-attacks and cyber security; Emerging trends and recent developments," Energy Reports, vol. 7, pp. 8176-8186, 2021.

[17] H.E. Yılmaz, A. Sirel, and M.F. Esen, "The impact of internet of things self-security on daily business and business continuity," in Research Anthology on Business Continuity and Navigating Times of Crisis, IGI Global, Hershey, PA, USA, pp. 695–712, 2022.

[18] M. Sarhan, S. Layeghy, N. Moustafa, M. Gallagher, and M. Portmann, "Feature extraction for machine learning-based intrusion detection in IoT networks," Digital Communications and Networks.

[19] M. Abdullahi, Y. Baashar, H. Alhussian, A. Alwadain, N. Aziz, L.F. Capretz, and S.J. Abdulkadir, "Detecting cybersecurity attacks in internet of things using artificial intelligence methods: A systematic literature review," Electronics, vol. 11, no. 2, p. 198, 2022.

[20] I. Ullah and Q. H. Mahmoud, "A scheme for generating a dataset for anomalous activity detection in IoT networks," in Canadian Conference on Artificial Intelligence, pp. 508–520, Springer, 2020.

[21] N. Moustafa and J. Slay, "UNSW-NB15: A comprehensive dataset for network intrusion detection systems (UNSW-NB15 network dataset)," in 2015 Military Communications and Information Systems Conference (MilCIS), pp.

[22] K. Dhanya, S. Vajipayajula, K. Srinivasan, A. Tibrewal, T.S. Kumar, and T.G. Kumar, "Detection of network attacks using machine learning and deep learning models," Procedia Computer Science, vol. 218, pp. 57–66, 2023.