Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 21 No. 6 (2025) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v21i06.53331

PAPER

Smart Defense: Harnessing Hybrid Deep Learning Models for Resilient IoT Intrusion Detection

Aouatif Arqane(✉), Omar Boutkhoum

LAROSERI Laboratory, Chouaib Doukkali University, El Jadida, Morocco

arqane.a@ucd.ac.ma

ABSTRACT

Internet of Things (IoT) networks have transformed various industries by enabling seamless connectivity and automation, yet they also pose significant security challenges. Traditional intrusion detection systems (IDS) struggle to protect these complex and diverse networks due to the vast variability in IoT devices, protocols, and communication patterns. This study explores the integration of deep learning (DL) and adversarial techniques to enhance IDS performance for IoT network security. We propose a DL-based IDS framework utilizing hybrid models, including convolutional neural network and long-short term memory (CNN-LSTM), bidirectional LSTM (B-LSTM), and bidirectional GRU (B-GRU). Experiments on the ToN-IoT dataset achieved accuracy levels exceeding 98% in non-adversarial scenarios. Among the models, B-LSTM exhibited outstanding resilience to adversarial attacks, such as FGSM, PGD, and Deep Fool, demonstrating its suitability for real-world IoT network security applications. This study highlights the need for robust IDS models to secure IoT networks effectively and emphasizes the importance of rigorous testing against adversarial threats, even when high accuracy is achieved.

KEYWORDS

Internet of Things (IoT), deep learning (DL), adversarial attacks, hybrid algorithm

1 INTRODUCTION

Internet of Things (IoT) networks, which drive automation and remote control across various sectors, also introduce significant security risks. Traditional intrusion detection systems (IDS) struggle to adapt to evolving threats, such as stealthy attacks and polymorphic malware, often failing to distinguish between normal and malicious activities across diverse devices and protocols. To enhance the security of IoT networks, deep learning (DL) presents promising solutions by utilizing neural networks to detect anomalies in these complex environments. DL-based IDS are more adaptable and resilient, effectively addressing the dynamic and ever-changing nature of IoT ecosystems. Moreover, incorporating adversarial attacks into IDS testing is crucial for assessing their robustness, as these attacks help evaluate how well

Arqane, A., Boutkhoum, O. (2025). Smart Defense: Harnessing Hybrid Deep Learning Models for Resilient IoT Intrusion Detection. *International Journal of Online and Biomedical Engineering (iJOE)*, 21(6), pp. 141–154. https://doi.org/10.3991/ijoe.v21i06.53331

Article submitted 2024-11-14. Revision uploaded 2025-01-16. Final acceptance 2025-02-21.

IDS can defend against sophisticated evasion techniques. This dual-testing approach provides valuable insights into the IDS's effectiveness in real-world scenarios.

This paper explores the integration of DL and adversarial techniques to strengthen the security of IoT networks. We propose a DL-based IDS framework using hybrid models, including convolutional neural network and long-short term memory (CNN-LSTM), bidirectional LSTM (B-LSTM), and bidirectional GRU (B-GRU). While other models, such as standalone LSTMs, GRUs, or transformers, have been applied in anomaly detection, our choice is motivated by the need to balance computational efficiency and performance for resource-constrained IoT environments. Standalone LSTM and GRU models are effective in capturing temporal dependencies; however, they fail to model spatial features inherent in IoT data streams generated by numerous sensors [1]. Transformers, on the other hand, excel in modeling long-range dependencies and handling complex sequences but come with significant computational overhead, making them less ideal for IoT devices with limited resources. In contrast, the hybrid models we propose offer distinct advantages. CNN-LSTM combines the spatial feature extraction capabilities of CNNs with the temporal modeling power of LSTMs, making it well-suited for detecting nuanced anomalies in IoT sensor data [2]. B-LSTM and B-GRU further enhance this temporal modeling by leveraging bi-directionality, which enables them to better capture dependencies in both past and future data points, an important consideration for IoT scenarios where sequential data integrity is key [3]. Furthermore, these selected models have demonstrated superior performance in handling continuous sensor data streams and detecting anomalies in various benchmarks. Their hybrid nature allows them to address the limitations of simpler architectures, such as LSTM, GRU, and CNN, which may struggle with capturing intricate relationships in time-series data [4].

To assess the resilience of our IDS, we rigorously tested it against adversarial attacks, including the Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and Deep Fool. These attacks simulate real-world intrusion attempts, providing valuable insights into the models' robustness. This evaluation confirms the effectiveness of our hybrid IDS framework in defending against various adversarial strategies, significantly improving the security of IoT networks.

This study makes several key contributions aimed at advancing IoT device security by developing resilient IDS models capable of countering emerging threats and adversarial attacks:

- Development of hybrid DL-based IDS models: we introduce three hybrid IDS models tailored to enhance IoT network security.
- Extensive experimental analysis: a comprehensive experimental analysis is conducted using the ToN-IoT dataset, incorporating preprocessing, data balancing, normalization, and hyperparameter tuning to ensure robust evaluation.
- Resilience against adversarial attacks: The models are tested against common adversarial attacks, FGSM, PGD, and Deep Fool, to evaluate their robustness.

The remainder of the paper is structured as follows: Section 2 reviews related studies on DL-based IDS and adversarial attacks targeting such systems. Section 3 describes the methodology. In Section 4, we explain the experimental setup, including the dataset, data processing techniques, hybrid models, and adversarial attack implementations. Section 5 presents the evaluation and analysis, identifying the optimal model. Finally, Section 6 offers conclusions, discusses limitations, and suggests future research directions.

2 RELATED STUDIES

2.1 DL-based IDS

[5] introduced a DL-based IDS tailored for IoT security, employing the SpiderMonkey Optimization algorithm (SMO) to bolster feature learning and the Stacked Deep Polynomial Network (SDPN) classifier. This approach showcased better accuracy and shorter training times than alternative methods, with evaluation of the NSL-KDD dataset yielding good results, including an accuracy of 99.02% and a precision of 99.38%. Similarly, [6] proposed a CNN-based method specifically designed for anomaly-based IDS in IoT networks. Their model was developed to analyze entire traffic flows within IoT environments and demonstrated robust performance, achieving accuracies of 99.51% and 95.55% on the NID and BoT-IoT datasets, respectively. Moreover, [7] presented a hybrid algorithm, DeepIoT, combining an enhanced Gaussian-Bernoulli restricted Boltzmann machine (DeepGB-RBM) for feature learning with a weighted deep neural network (WDNN) classifier. Evaluation of the CICIDS2017 dataset exhibited the superior performance of DeepIoT IDS, achieving detection accuracies of 99.38% and 99.99% for web and bot attack scenarios.

2.2 Attacking DL-based IDS

The research conducted by [8] introduces a novel method for generating adversarial network packets targeting modern DL-based NIDSs. Their approach utilizes model extraction and saliency mapping techniques, resulting in successful attacks on Kitsune, a prominent NIDS. These attacks led to an increase in false positives and false negatives in scenarios involving the Mirai botnet and video streaming, emphasizing the significant impact of minor alterations to network packets on NIDS prediction accuracy. [9] focused on evaluating the effectiveness of evasion attacks and strategies to enhance the resilience of DL-based IDSs. They explore the use of ANN, CNN, and Recurrent Neural Networks (RNN) models, employing a min-max approach to train robust IDSs against adversarial examples. Evaluation results demonstrate that adversarial training through this approach improves the IDS's robustness against various adversarial techniques. [10] investigated the vulnerability of DL-NIDSs, particularly Kitsune, against adversarial attacks. Their experiments highlight the need to evaluate DL-NIDSs against adversarial machine learning techniques, as attackers may exploit human knowledge utilized in current DL-NIDSs to increase the likelihood of successful attacks. This underscores the importance of continually assessing DL-NIDS's resilience to emerging adversarial threats as the field evolves.

2.3 Gaps and limitations

The analysis of reviewed studies indicates that DL-based IDS for IoT network security remains underexplored, emphasizing the need for further research. While many studies focus on enhancing DL models, they often neglect the selection of relevant, up-to-date datasets, which are critical for accurate evaluations. Key factors such as data preparation, dimensionality reduction, and hyperparameter tuning receive insufficient attention, impacting model effectiveness. Furthermore, reliance on single DL models limits the ability to capture sequential temporal patterns crucial in IoT network traffic. Standard metrics such as accuracy and precision are widely used, yet other essential

evaluation metrics are frequently overlooked. Additionally, few studies rigorously test their models against adversarial attack scenarios, missing valuable insights into vulnerabilities. This underscores the need for comprehensive methodologies to bridge these gaps and improve the effectiveness of DL-based IDS in IoT network security.

3 METHODOLOGY OVERVIEW

The proposed framework's methodology begins with acquiring data from the ToN-IoT dataset, followed by thorough data preprocessing. The dataset is divided into training and testing sets, with 70% allocated for training and 30% for testing. During the deep learning training phase, three distinct models that are CNN-LSTM, BLSTM, and B-GRU are trained using the training data. Bayesian optimization (BO) is employed to fine-tune the hyperparameters of these models, aiming to optimize their performance. The optimal hyperparameters are selected based on maximizing the validation accuracy of the models. Once identified, the models are trained using the optimized hyperparameters. Following the training phase, the models undergo evaluation using the testing data. Various evaluation metrics are employed to assess the performance of each model. Upon completing the performance evaluation, the methodology progresses to the robustness evaluation phase, where three different attack techniques, FGSM, PGD, and DeepFool, are executed on the trained models. The robustness of each model against these attacks is evaluated using a set of evaluation metrics. Finally, the best-performing model is selected based on its effectiveness during the evaluation phases. This optimal model is deployed as an IDS to identify and classify normal and attack anomalies within the IoT network. Figure 1 illustrates the overall workflow of the proposed framework.

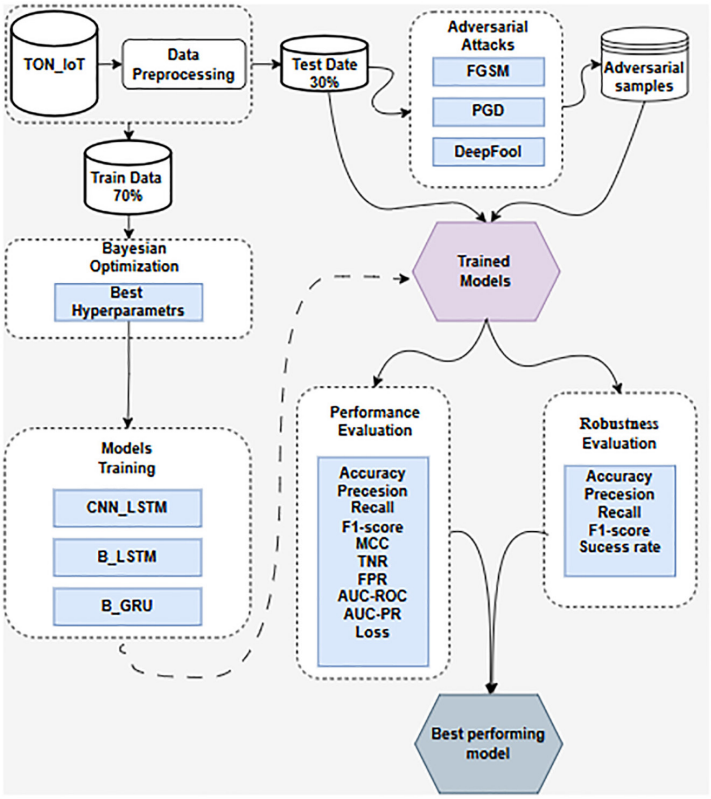

Fig. 1. Workflow of the proposed framework

4 EXPERIMENTAL SETUP

4.1 Dataset description

The ToN-IoT dataset [11] stands as a comprehensive repository of real-world Industry 4.0/IoT data, meticulously gathered from a systematic testbed established at UNSW Canberra Cyber IoT Lab, School of Engineering and Information Technology (SEIT), UNSW Canberra at The Australian Defense Force Academy (ADFA). For the current study, the Network dataset from ToN-IoT is employed to illustrate the IoT devices scenario. The Train Test Network dataset utilized is the officially released subset of the ToN-IoT dataset. This subset comprises a total of 461,043 records, with 300,000 labeled as 'normal' and 161,043 labeled as 'anomalous.' Each attack class consists of 20,000 flows, except for XSS attacks, which are represented by 1,043 flow records. Additionally, for IoT attack classification, 43 features are extracted and categorized into subsets based on the type of information they encapsulate, encompassing statistical activity features, connection activity features, HTTP activity features, violation activity features, SSL activity features, and DNS activity features.

4.2 Data processing

Before feeding data into DL algorithms, data cleaning and preparation are essential for optimal performance. During our experimentation, we faced challenges such as categorical features, class imbalance, and irrelevant attributes that could affect the model's effectiveness. To address these issues, we applied various preprocessing and normalization techniques to improve data quality and ensure better results.

- Data cleaning and preparation: First, irrelevant features such as 'source port,' 'timestamp,' 'destination port,' and 'IP addresses' were removed to prevent overfitting during model training. Next, duplicate records were eliminated to preserve data integrity and reduce bias in the analysis. Additionally, categorical features were transformed into numerical values using 'one-hot encoding,' resulting in a final dataset with 127 features.
- Data normalization: The dataset contained many outliers and heterogeneous records, which could skew model predictions. To address this, the RobustScaler normalization technique was applied. Unlike standard methods, RobustScaler uses the median and Interquartile Range (IQR) to rescale data, making it less sensitive to outliers and ensuring more reliable feature scaling.
- Data balancing: The ToN-IoT dataset faced a class imbalance issue, which was addressed by exploring oversampling, undersampling, and a hybrid approach. To avoid the risks of overfitting from oversampling and losing crucial information from undersampling the hybrid SMOTE method was chosen. SMOTE generates new minority class instances and selectively removes some majority class instances. Before balancing, the dataset had 117,045 normal instances and 69,810 anomalous ones; after applying SMOTE, both classes were balanced with 117,045 instances each.

4.3 Proposed hybrid models

Hyperparameter tuning. Hyperparameters play a crucial role in DL models by balancing performance, memory, and computational complexity. Parameters such as learning rate, iteration count, hidden layers, and batch size need tuning to optimize accuracy and minimize loss. In this study, BO is employed for efficient hyperparameter tuning, providing systematic improvements with fewer computational resources by using probabilistic models to guide the search and refine resource allocation. Table 1 summarizes the results of BO for optimizing the hyperparameters of the proposed hybrid DL models.

Table 1. Optimal model parameters by Bayesian optimization

Hyperparameters	Range Values	Best Values		
		CNN-LSTM	B-LSTM	B-GRU
Number of units	32 to 256, step: 32	256	32	64/64
Dense layer size	16 to 128, step: 16	90/80	64	64
Activation function	Relu, Sogmoid	Relu	Relu	Relu
Dropout size	0.2 to 0.7, step: 0.1	0.6	0.6	0.3
Optimizer	Adam, SGD, RMSprop	Adam	Adam	Adam
Learning rate	0.0001 to 0.005, step: 0.001	0.001	0.002	0.001

CNN-LSTM. The CNN-LSTM hybrid model leverages the strengths of CNNs in feature extraction and LSTMs in capturing temporal dependencies. The model has a 1D convolutional layer with 256 filters, a kernel size of 5, and a ReLU activation function to perform feature extraction. Followed by a MaxPooling layer with a pool size of 2 to reduce the spatial dimensions of the features. A Dropout layer with a rate of 0.62 is applied to mitigate overfitting. Next, two LSTM layers with 90 units each are employed to capture sequential patterns in the data. Another dropout layer is added for regularization. Finally, a Dense layer with a single unit and a Sigmoid activation function is used for binary classification. The advantage of the proposed CNN_LSTM-based IDS IoT networks lies in their ability to effectively capture and classify complex patterns in network traffic data. By combining convolutional and recurrent layers, the model can efficiently extract spatial and temporal features from network traffic data, which is crucial for detecting various types of intrusions and anomalies in IoT networks. Additionally, the incorporation of dropout layers helps prevent overfitting and enhances the model's generalization ability, leading to improved performance and robustness.

Bidirectional LSTM. The architecture of the bidirectional LSTM (BLSTM) model comprises several key components carefully designed to accurately detect intrusions within IoT environments while maintaining robustness and generalization capability. Beginning with an input layer configured to accommodate variable-length sequences of features. Also, the model incorporates a bidirectional LSTM layer with 32 units that enables the network to process input sequences in both forward and backward directions. This is crucial for detecting behavior patterns over time in IoT environments, where activities and interactions may unfold gradually. Additionally, an attention mechanism is employed to dynamically weigh the importance of different parts of the input sequence, enabling the model to focus on relevant information

while disregarding irrelevant noise. The 'use scale' parameter is set to 'True,' which scales the attention scores to improve stability during training. The output of the bidirectional LSTM layer and the attention scores are concatenated and globally pooled to create a fixed-length representation of the input data. This facilitates the identification of anomalous behavior and potential security threats. Also, a Global Max Pooling Layer performs global max pooling on the concatenated output, generating a fixed-length representation of the data that captures the most salient features across the entire sequence. After that, dropout regularization and batch normalization layers help prevent overfitting by introducing randomness during training and stabilizing the learning process. This ensures that the model generalizes well to unseen data and maintains robust performance in real-world scenarios. Finally, we have two dense layers with ReLU activation functions. The first dense layer has 64 units and serves as a feature extractor. In contrast, the second dense layer with a single unit and sigmoid activation function performs binary classification, predicting whether an intrusion is detected.

Bidirectional GRU. The third model utilizes bidirectional GRU layers to effectively handle sequential data, a common occurrence in IoT environments where sensor readings and device interactions unfold over time. This model architecture enables the capture of temporal patterns and dependencies crucial for anomaly detection or intrusion identification. Bidirectional GRU layers process input sequences both forwards and backward, incorporating context from past and future time steps. This bidirectional processing enhances the model's understanding of the contextual nuances surrounding each data point, thereby improving its ability to discern anomalous behavior in IoT network activities. Two bidirectional GRU layers were used with 64 units each and the ReLU function. Two dropout layers are strategically inserted after each Bidirectional GRU layer, with rates of 0.3 and 0.4, respectively, to curb overfitting, preventing the model from memorizing noise or irrelevant patterns in the training data and enhancing its generalization on unseen data. Additionally, batch normalization is employed to normalize activations from the preceding layer, expediting the training process and enhancing stability. Following the GRU layers, the Dense layer was included with 64 units and the ReLU function to perform nonlinear transformations on learned features, allowing the model to capture intricate relationships and patterns in the input data, thus enhancing its ability to distinguish between normal and abnormal behaviors in IoT networks. Lastly, the output layer comprises a single neuron outputting a probability value indicating the likelihood of the input belonging to the normal class in the binary classification task.

4.4 Adversarial attacks

The attacks, including FGSM, PGD, and DeepFool, were implemented from scratch using Python libraries, with TensorFlow handling gradient computation and tensor manipulation. Creating these attacks from scratch allowed for customization and fine-tuning to match specific model architectures and datasets, enhancing both flexibility and understanding of the attacks. For these attacks, parameter values were selected after extensive experimentation with various settings. In FGSM, ε (epsilon) was set to 0.1 to balance perturbation strength with minimal perceptibility. For PGD, $\varepsilon = 0.1$ controlled the maximum perturbation, and α (alpha) = 0.01 was chosen to refine step sizes over 20 iterations, optimizing both effectiveness and computational efficiency.

In DeepFool, $\varepsilon = 0.1$, and 20 iterations were chosen after testing various values to maintain strong perturbations while remaining computationally feasible. These values, derived from multiple trials, were consistently applied across attacks for fair and meaningful comparisons. Table 2 presents the parameters utilized in the employed attacks.

Table 2. Attack's parameters

Attack	Parameters
FGSM	$\varepsilon = 0.1$
PGD	$\varepsilon = 0.1$, $\alpha = 0.01$, num iter = 20
DeepFool	$\varepsilon = 0.1$, input data, num classes = 2, max iter = 20

4.5 Evaluation metrics

To comprehensively evaluate the effectiveness of the proposed DL models, we employed a diverse set of evaluation metrics, encompassing accuracy (acc), precision, recall, F1-score, Matthews correlation coefficient (MCC), TNR, FPR, AUC-ROC, AUC-PR, and loss function. These metrics were carefully chosen to offer a thorough description of the outcomes obtained from DL-based IDS using the ToN-IoT dataset. Metrics such as the F1 Score and MCC illustrate the model's ability to handle class imbalances effectively, highlighting superiority over traditional metrics such as accuracy. The inclusion of TNR and FPR allows us to demonstrate that our models not only detect attacks with high accuracy but also maintain low false positive rates, ensuring smooth network operation. Metrics such as AUC-ROC and AUC-PR, especially when combined with adversarial attack scenarios, emphasize the models' ability to consistently distinguish between attack and benign traffic under challenging conditions. This comprehensive evaluation aligns with our goal of developing IDS models tailored for real-world IoT applications. The use of these metrics substantiates the claim that our models excel in detecting attacks, reducing false alarms, and maintaining resilience in adversarial settings.

5 EVALUATION AND ANALYSIS

5.1 Performance evaluation

To ensure a consistent comparison of the proposed models, we trained them under identical conditions. Thus, we fixed the batch size at 64 and conducted training for 10 epochs. Employing the 'early stopping' function allowed us to identify the optimal number of epochs, with observations indicating that accuracy and loss stabilization occurred across nearly all models after 10 epochs.

According to the results displayed in Table 3, all models demonstrate high accuracy, with B-GRU achieving the highest at 98.58%, followed closely by B-LSTM at 98.52% and CNN-LSTM at 98.00%. Precision, recall, and F1-score metrics also exhibit strong values across all models, indicating robust performance in both malicious and normal class predictions. B-GRU shows slightly superior precision, recall, and F1-score compared to the others, suggesting a better balance between identifying positive instances and minimizing false positives. Additionally, B-GRU boasts the

highest MCC, signifying superior overall performance in capturing underlying data patterns and making accurate predictions. Also, all models demonstrate minimal FPR values, showcasing their effectiveness in avoiding unnecessary alerts. B-GRU exhibits a slightly higher TNR, indicating its enhanced ability to correctly classify normal activities. Moreover, all models exhibit high values for AUC-ROC and AUC-PR, highlighting robust performance.

Table 3. Performance results of the three models

Model	Acc (%)	Precision	Recall	F1-Score	MCC	TNR	FPR	AUCROC	AUC-PR
CNN-LSTM	98.00	0.973	0.987	0.980	0.960	0.972	0.027	0.980	0.983
B-LSTM	98.52	0.980	0.990	0.985	0.970	0.980	0.019	0.985	0.987
B-GRU	98.58	0.980	0.991	0.986	0.971	0.980	0.019	0.985	0.988

Upon comparing the confusion matrices of the three models, distinct patterns emerge. All models showcase a high number of TP and TN, underscoring their proficiency in accurately classifying both normal and anomalous instances. However, B-GRU and B-LSTM exhibit fewer FPs and FNs in comparison to CNN-LSTM. This implies that B-GRU and B-LSTM demonstrate slightly enhanced efficacy in averting false alarms and mitigating misclassifications of normal instances as anomalous. Moreover, B-GRU and B-LSTM demonstrate superior balance in curtailing FP and FN relative to CNN-LSTM, indicating their heightened reliability in discerning between normal and anomalous activities within IoT environments. Consequently, the marginally superior performance of B-GRU and B-LSTM in minimizing FP and FN compared to CNNLSTM suggests their suitability for intrusion detection. Figure 2 illustrates the confusion matrix of the three models.

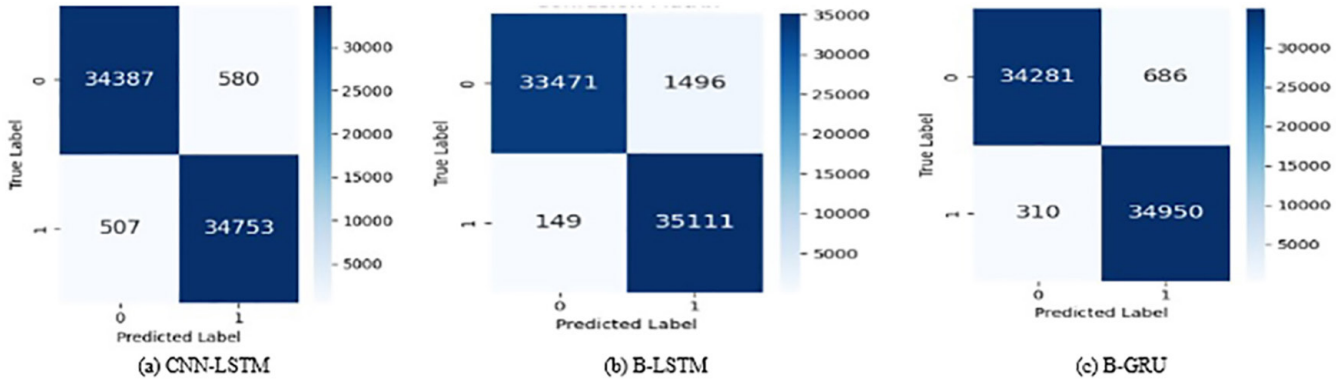

Fig. 2. Confusion matrix of the three models

Based on the comparison of accuracy and loss graphs depicted in Figures 3 and 4, the CNN-LSTM model shows a gradual increase in accuracy, reaching approximately 97.23% on training data and 98.05% on validation data by the end of training, with validation loss decreasing to around 0.0664. The B-LSTM model also improves accuracy over epochs, attaining about 98.09% on training data and 98.58% on validation data, with validation loss dropping to approximately 0.0457. The B-GRU model, while achieving an accuracy of around 98.42% on training data and 98.58% on validation data, shows slight fluctuations in loss, settling at around 0.1192. The CNN-LSTM and B-LSTM models exhibit smoother convergence in loss compared

to B-GRU, with B-LSTM converging faster and achieving similar accuracy in fewer epochs. In summary, all three models demonstrate promising performance for intrusion detection in IoT networks, with B-LSTM exhibiting slightly faster convergence and comparable accuracy to B-GRU.

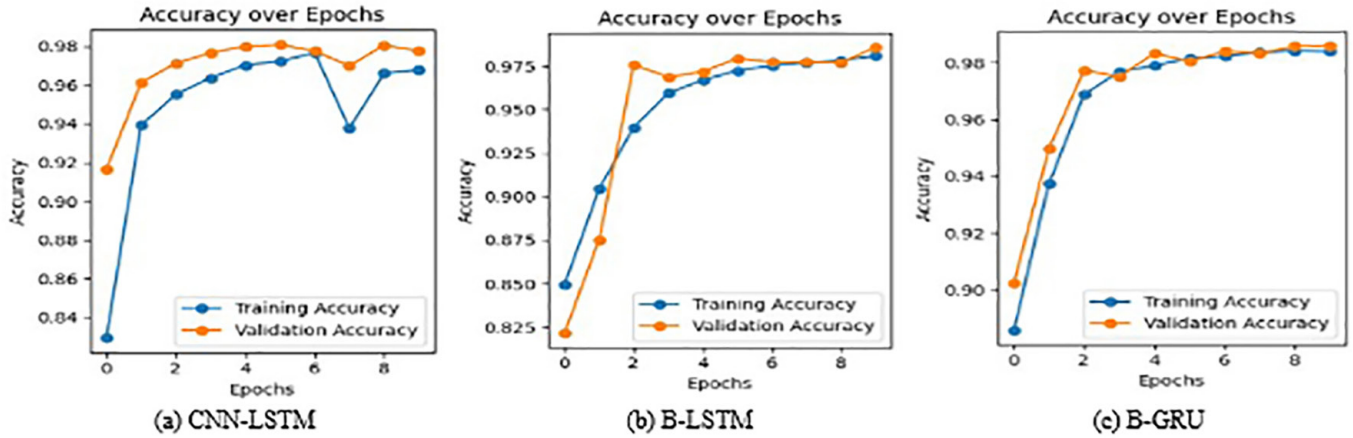

Fig. 3. Accuracy of the three models over epochs

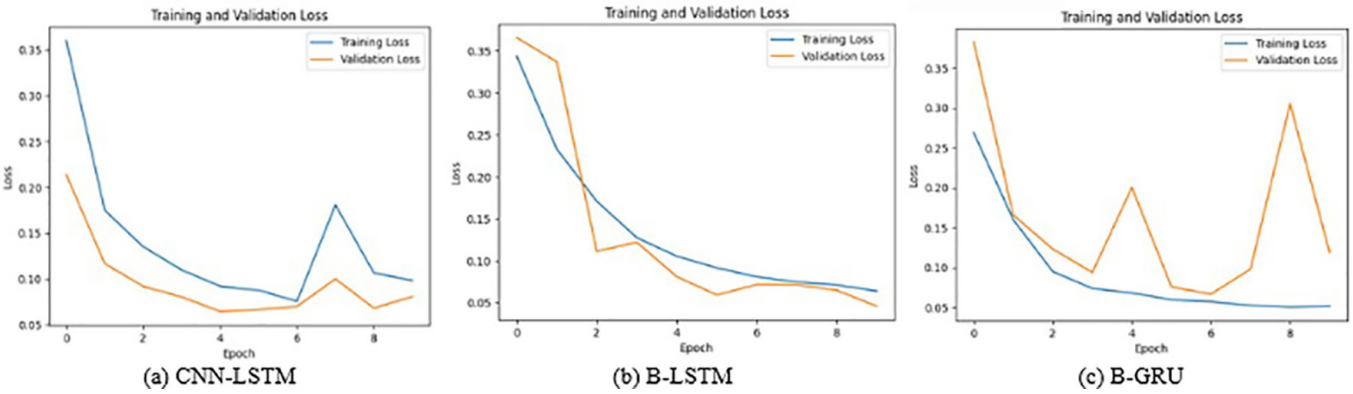

Fig. 4. Loss of the three models over epochs

5.2 Robustness evaluation

In this study, we conducted a comprehensive evaluation of the robustness of our proposed models against adversarial attacks (FGSM, PGD, and DeepFool) to ensure their reliability and security in practical applications.

Table 4 highlights each model's performance, revealing strengths and vulnerabilities across various attack scenarios to guide the selection of reliable architectures for deployment. Starting with the CNN-LSTM model, it shows high accuracy under FGSM 66.35% and PGD 63.30% attacks but suffers under DeepFool, with accuracy dropping to 51.20%. Despite high recall with DeepFool 0.997, its overall precision and F1-score are lower, indicating a higher FPR. This model's vulnerability to DeepFool is evident, with a success rate of 48.79%. Moving to the B-LSTM model, it surpasses CNN-LSTM in accuracy under FGSM 70.98% and has the highest recall and F1-score for this attack, showing a good balance between precision and recall. However, it also experiences a significant accuracy drop to 61.13% under DeepFool, with a success rate of 38.69%. Despite this, B-LSTM offers better resistance

to DeepFool compared to CNN-LSTM. Lastly, the B-GRU model performs moderately across all attacks, with accuracy ranging from 54.89% to 63.15%. It has notable precision under PGD 0.818 but the lowest accuracy and F1-score under DeepFool, with a high success rate of 61.87%. This indicates significant vulnerability to DeepFool, suggesting B-GRU's performance may be less reliable in practical applications.

Table 4. Performance of the models under the adversarial attacks

Model	Attack	Acc (%)	Precision	Recall	F1-Score	Success Rate (%)
CNN-LSTM	FGSM	66.35	0.641	0.716	0.677	33.64
	PGD	63.30	0.648	0.605	0.626	36.69
	DeepFool	51.20	0.510	0.997	0.674	48.79
B-LSTM	FGSM	70.98	0.636	0.846	0.726	29.01
	PGD	61.14	0.724	0.246	0.368	38.54
	DeepFool	61.13	0.544	0.925	0.685	38.69
B-GRU	FGSM	54.89	0.541	0.553	0.547	45.10
	PGD	63.15	0.818	0.352	0.493	36.84
	DeepFool	38.12	0.419	0.565	0.481	61.87

Figure 5 depicts the accuracy of the three models under adversarial-free conditions and the adversarial attacks: FGSM, PGD, and DeepFool. In the absence of adversarial manipulation, all models demonstrate high accuracy levels exceeding 98%. However, when subjected to attacks, there is a noticeable decline in accuracy across all models. B-LSTM consistently exhibits superior resilience against adversarial attacks, outperforming the other models in most scenarios. Notably, under the FGSM attack, B-LSTM achieves the highest accuracy of 70.98%, followed closely by CNNLSTM. Conversely, under the DeepFool attack, CNN-LSTM demonstrates the highest accuracy among the models, although all models suffer a significant decrease in performance.

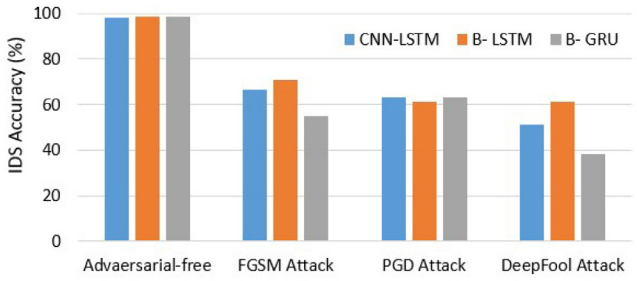

Fig. 5. Accuracy of models under attacks

Choice of the optimal model: Based on the comprehensive analysis of model performance under various adversarial attacks, the BLSTM model emerges as the most robust choice among CNN-LSTM, B-LSTM, and B-GRU. Despite CNN-LSTM demonstrating high accuracy under benign conditions, its susceptibility to the DeepFool attack, as indicated by a success rate of 48.79%, underscores its vulnerability to sophisticated adversarial manipulations. Conversely, the B-LSTM model showcases resilience against the DeepFool attack, with a lower success rate of

38.69%, suggesting better resistance to subtle perturbations introduced by DeepFool. Moreover, B-LSTM outperforms CNN-LSTM in terms of accuracy under the FGSM attack, achieving a value of 70.98%, and exhibits the highest recall and F1-score among all models under this attack, indicating its robustness in correctly identifying instances of the target class while maintaining a balance between precision and recall. Additionally, the graph depicting model accuracy under adversarial-free conditions and various attacks further highlights B-LSTM's consistent superiority in resilience against adversarial manipulations, reinforcing its suitability for intrusion detection tasks in IoT environments. Therefore, based on the analysis of attack scenarios and model performance, the B-LSTM model emerges as the optimal choice for ensuring robust and reliable intrusion detection in real-world applications.

5.3 Comparative analysis

We perform a comparative analysis of studies utilizing the ToN-IoT dataset, evaluating various models such as LSTM, XGBoost, RandomForest, and DenseNet, which achieve accuracies between 96.56% and 99.1%. Our B-LSTM framework demonstrates a notable accuracy of 98.52% under normal conditions and maintains a robust performance with 70.98% accuracy even during adversarial attacks. In comparison, [13] reports a DNN model achieving 87.0% accuracy normally but only 48.0% under attack, highlighting our framework's superior resilience. While the results indicate high performance across models, it is essential to assess their robustness against adversarial conditions. Table 5 provides a comprehensive comparison of model accuracies across different studies.

Table 5. Comparison of model accuracies in a few studies

Study	Model	Acc (%)	Acc Under Attacks (%)
[14]	LSTM	96.56	–
[15]	XG-Boost	99.1	–
[16]	Random Forest	98.39	–
[17]	LSTM	97.5	–
[18]	DenseNet	98.5	–
[13]	DNN	87.0	48.0
Our framework	B-LSTM	98.52	70.98

6 CONCLUSION

In this study, we explored the integration of hybrid DL and adversarial techniques to enhance IoT network security. Using the ToN-IoT dataset, all models achieved over 98% accuracy, demonstrating the effectiveness of our DL-based IDS framework. The B-GRU model had the highest F1 score, precision, and recall, indicating optimal balance in minimizing FP and FN. B-LSTM demonstrated the fastest convergence and comparable performance, making it ideal for rapid deployment in practical settings. Both models performed excellently in minimizing FP, while all models showed high AUC-ROC and AUC-PR, indicating strong class discrimination capabilities under adversarial conditions. B-LSTM proved to be the most resilient, maintaining high

accuracy along with strong precision, recall, and F1-score values, thereby ensuring effective classification of both normal and malicious activities. Its robust performance across multiple metrics, particularly in minimizing FP and FN, demonstrates its reliability in adversarial environments, making it the optimal model for practical, secure intrusion detection in IoT networks.

Limitations and future work: Like any research endeavor, our study is subject to certain limitations. One limitation is the focus solely on a specific set of adversarial attacks. So, future research should explore a broader range of strategies to assess DL-based IDS resilience in IoT networks better. The evaluation used a single dataset; future work should test the models with diverse datasets to evaluate their generalization.

7 REFERENCES

[1] R. Z. Khalid, A. Ullah, A. Khan, A. Khan, and M. H. Inayat, "Comparison of standalone and hybrid machine learning models for prediction of critical heat flux in vertical tubes," *Energies*, vol. 16, no. 7, p. 3182, 2023. https://doi.org/10.3390/en16073182

[2] H. Alkahtani and T. H. H. Aldhyani, "Botnet attack detection by using CNN-LSTM model for Internet of Things applications," *Security and Communication Networks*, vol. 2021, no. 1, p. 3806459, 2021. https://doi.org/10.1155/2021/3806459

[3] F. M. Shiri, T. Perumal, N. Mustapha, and R. Mohamed, "A comprehensive overview and comparative analysis on deep learning models: CNN, RNN, LSTM, GRU," *arXiv preprint arXiv:2305.17473,* 2024. https://doi.org/10.48550/arXiv.2305.17473

[4] K. P. Rasheed Abdul Haq and V. P. Harigovindan, "Water quality prediction for smart aquaculture using hybrid deep learning models," *IEEE Access*, vol. 10, pp. 60078–60098, 2022. https://doi.org/10.1109/ACCESS.2022.3180482

[5] Y. Otoum, D. Liu, and A. Nayak, "DL-IDS: A deep learning–based intrusion detection framework for securing IoT," *Transactions on Emerging Telecommunications Technologies*, vol. 33, no. 3, p. e3803, 2022. https://doi.org/10.1002/ett.3803

[6] T. Saba, A. Rehman, T. Sadad, H. Kolivand, and S. A. Bahaj, "Anomaly-based intrusion detection system for IoT networks through deep learning model," *Computers and Electrical Engineering*, vol. 99, p. 107810, 2022. https://doi.org/10.1016/j.compeleceng.2022.107810

[7] Z. Maseer, R. Yusof, S. Mostafa, N. Bahaman, O. Musa, and B. Ali, "DeepIoT.IDS: Hybrid deep learning for enhancing IoT network intrusion detection," *CMC*, vol. 69, no. 3, pp. 3945–3966, 2021. https://doi.org/10.32604/cmc.2021.016074

[8] H. Qiu, T. Dong, T. Zhang, J. Lu, G. Memmi, and M. Qiu, "Adversarial attacks against network intrusion detection in IoT systems," *IEEE Internet of Things Journal*, vol. 8, no. 13, pp. 10327–10335, 2021. https://doi.org/10.1109/JIOT.2020.3048038

[9] R. A. Khamis and A. Matrawy, "Evaluation of adversarial training on different types of neural networks in deep learning-based IDSs," in *2020 International Symposium on Networks, Computers and Communications (ISNCC)*, 2020, pp. 1–6. https://doi.org/10.1109/ISNCC49221.2020.9297344

[10] J. Clements, Y. Yang, A. A. Sharma, H. Hu, and Y. Lao, "Rallying adversarial techniques against deep learning for network security," in *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2021, pp. 1–8. https://doi.org/10.1109/SSCI50451.2021.9660011

[11] "TON_IoT datasets–OneDrive." Consulté le: 24 mars 2024. [En ligne]. Disponible sur: https://unsw-my.sharepoint.com/personal/z5025758_ad_unsw_edu_au/_layouts/15/onedrive.aspx?id=%2Fpersonal%2Fz5025758%5Fad%5Funsw%5Fedu%5Fau%2FDocuments%2FTON%5FIoT%20datasets&ga=1

[12] H. Cho, Y. Kim, E. Lee, D. Choi, Y. Lee, and W. Rhee, "Basic enhancement strategies when using bayesian optimization for hyperparameter tuning of deep neural networks," *IEEE Access*, vol. 8, pp. 52588–52608, 2020. https://doi.org/10.1109/ACCESS.2020.2981072

[13] E. Lella, N. Macchiarulo, A. Pazienza, D. Lofù, A. Abbatecola, and P. Noviello, "Improving the robustness of DNNs-based network intrusion detection systems through adversarial training," in *2023 8th International Conference on Smart and Sustainable Technologies (SpliTech)*, 2023, pp. 1–6. https://doi.org/10.23919/SpliTech58164.2023.10193009

[14] R. A. Elsayed, R. A. Hamada, M. I. Abdalla, and S. A. Elsaid, "Securing IoT and SDN systems using deep-learning based automatic intrusion detection," *Ain Shams Engineering Journal*, vol. 14, no. 10, p. 102211, 2023. https://doi.org/10.1016/j.asej.2023.102211

[15] A. Gad, M. Haggag, A. Nashat, and T. Barakat, "A distributed intrusion detection system using machine learning for IoT based on ToN-IoT dataset," *International Journal of Advanced Computer Science and Applications*, vol. 13, pp. 548–563, 2022. https://doi.org/10.14569/IJACSA.2022.0130667

[16] M. Wang, N. Yang, and N. Weng, "Securing a smart home with a transformer-based IoT intrusion detection system," *Electronics*, vol. 12, no. 9, 2023. https://doi.org/10.3390/electronics12092100

[17] A. Ahmim, L. Maglaras, M. A. Ferrag, M. Derdour, and H. Janicke, "A novel hierarchical intrusion detection system based on decision tree and rules-based models," in *2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, 2019, pp. 228–233. https://doi.org/10.1109/DCOSS.2019.00059

[18] I. Tareq, B. M. Elbagoury, S. El-Regaily, and E.-S. M. El-Horbaty, "Analysis of ToN-IoT, UNW-NB15, and Edge-IIoT datasets using DL in cybersecurity for IoT," *Applied Sciences*, vol. 12, no. 19, 2022. https://doi.org/10.3390/app12199572

8 AUTHORS

Aouatif Arqane is currently pursuing a Ph.D. degree in Machine Learning at the Department of Computer Science, Chouaib Doukkali University, El Jadida, Morocco. Her research interests include the application of machine learning and deep learning in cybersecurity (E-mail: arqane.a@ucd.ac.ma).

Omar Boutkhoum is an Associate Professor at the Computer Science department in the Faculty of Sciences of Chouaib Doukkali University, EL Jadida, Morocco. He received his Ph.D. degree in Computer Science from the Faculty of Sciences and Techniques of Caddi Ayyad University, Marrakesh, in 2017. His research interests are in the application of decision support systems, knowledge graphs, and artificial intelligence (E-mail: boutkhoum.o@ucd.ac.ma).