# Adversarially-Resilient Deep Learning Framework for Detecting Evasive Cyber Threats in Network Traffic

Nasser Alsharif
Department of Sciences and Technology, Ranyah University Collage,
Taif University, Saudi Arabia
n.alsharif@tu.edu.sa

***Abstract:*** *Standard Intrusion Detection Systems (IDS) have become increasingly inadequate due to the rampant sophistication of cyberattacks (ranging from stealthy advanced persistent threats, to polymorphic malware). Leading-edge models based on deep learning have recognized the power of learning complex spatial, and temporal, patterns from inputs and can learn anomalous or evasive threats, but they are vulnerable to adversarial examples (inputs that are slightly modified to cause misclassifications). This work introduces an adversarially-resilient deep learning framework for detecting evasive (or stealthy) cyber threats in the context of network traffic. The framework blends One-Dimensional Convolutional Neural Networks (1D-CNNs) with Long Short-Term Memory (LSTM) layers in the same architecture to model spatial correlations along with the temporal sequence of flow-based network features. Efficient feature selection is performed using Recursive Feature Elimination (RFE) to eliminate a large amount of noise and dimensionality. Finally, adversarial training utilizing the Fast Gradient Sign Method (FGSM) is incorporated to create ferry adversaries to the model. The construct is trained and tested with CICIDS2017 to validate the methodology, benchmarking a number of modern day attack scenarios in the process. Experiment-based results reveal the superior performance of the proposed framework, achieving a clean test data F1-score of 91.7%, and a robust performance above the 80% threshold with an F1-score of 84.5% under FGSM-based attacks. Tests comparing the proposed model, and its robustness and accuracy under similar attacks against the baseline models Support Vector Machine (SVM), Random Forest and vanilla Multilayer Perceptron (MLP) substantiate the advantage of the system we proposed on both points Robustness, and Accuracy. These results suggest the use of adversarially trained hybrid deep learning models for intrusion detection in real-world applications is worth pursuing, providing an effective and adaptable way to secure critical infrastructure in the digital era.*

***Keywords:*** *adversarial, deep Learning, IDS, network traffic, cyber threats*.

*Received July 10, 2025*; *Accepted December 08, 2025*
*https://doi.org/10.34028/iajit/23/4/11*

## 1. Introduction

The increasing interconnection among computer systems has transformed the world, promoting innovation, ease of use, and scalability in areas such as finance, healthcare, education, and national defense. The technology has also brought digital infrastructures face to face with unprecedented cyber threats. Modern cyberattacks, ranging from data breaches and ransomware to those Advanced Persistent Threats (APTs), are stealthy, persistent, and evasive [8]. Such attacks also tend to employ evasion techniques such as polymorphism, obfuscation, and encryption to evade normal defense techniques, rendering most legacy security techniques incapable of detecting or delaying new and upcoming threats [14].

Legacy Intrusion Detection Systems (IDS) are heavily dependent on preconfigured signatures or rule-based manual engineering to detect malicious behavior [12]. Although they are effective in detecting known attacks, they cannot identify new or obfuscated attacks, particularly those that make slight alterations to stay under the radar. In addition machine learning which has been used as a better alternative to rule-based systems, cannot generalize well across wide-ranging types of attacks and tend to need a lot of feature engineering and domain expertise [10].

Deep learning has become an effective means of cybersecurity because it can learn high-level, hierarchical data representations with minimal human effort. Deep learning models like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been used to identify intrusions by learning spatial and temporal patterns of the network traffic [6]. These models are capable of automatically learning latent patterns and interactions among network flow features that would be hard to learn through traditional methods, leading to significantly enhanced detection capability across a broad range of threat types [2].

Although deep learning-based IDS have their strength, they are also vulnerable to adversarial attacks. Such attacks involve an attacker injecting input data with tiny, well-designed perturbations, which are mostly

imperceptible to humans but drive the model to make erroneous predictions [4]. For IDS, traffic samples that have been adversarially altered can be designed to strongly mimic normal network traffic and result in malicious packets being misclassified as normal. This weakness significantly hurts the credibility and safety of deep learning-based detection systems, particu larly those in life-critical use [15].

The root cause of the problem is the very non-linear decision boundaries of deep learning models, which make them sensitive to variations in the input data. This makes model behavior controllable by the attackers through manipulation of gradient information or exploitation of blind spots in the feature space learned [3]. Also, most deep learning models are "black boxes" with minimal interpretability, and it becomes difficult for analysts to be aware of the assumptions of their predictions or when a model has been manipulated. To counter these weaknesses, this study proposes an adversarially-resilient deep learning model for network traffic intrusion detection.

The objective is to obtain a robust model that not only performs with good accuracy in normal circumstances but also maintains high detection accuracy against adversarial inputs. The proposed framework leverages a hybrid model that employs one-Dimensional Convolutional Neural Networks (1D-CNNs) and Long Short-Term Memory (LSTM) layers to learn both spatial feature interactions and temporal patterns in network flow data. The convolutional layers capture local traffic feature patterns, and the LSTM units identify long-range sequential patterns that are generally characteristic of coordinated attacks. In simplifying and generalizing, the model employs Recursive Feature Elimination (RFE) from a Random Forest classifier to choose the most informative features out of the high-dimensional input space. This not only improves computational efficiency but also improves classification performance by eliminating noisy and redundant information.

To further make the model adversarial attack robust, adversarial training is incorporated into the training. Adversarial examples are generated using the Fast Gradient Sign Method (FGSM) which deforms input samples in the direction of loss gradient in order to mislead the model. The adversarial samples are combined with clean training data so that the model learns decision boundaries that are resistant. this work fills an essential lacuna in adversarial robustness of deep learning-based cybersecurity through the development of a model that can resist adversarial attacks. The adversarial training, optimally selected features, and the hybrid neural architecture provide an end-to-end solution to contemporary intrusion detection problems. By improving the resilience and dependability of deep learning frameworks, this research helps in the creation of safe, scalable, and smart systems that can protect against advanced cyber threats.

## 2. Related Work

Research into the creation of adversarially-resilient deep learning justifications for the detection of evasive cyber threats in network traffic is a burgeoning field that uses complex architectures based on advanced neural networks, and unique training approaches. A number of studeis show how deep learning can be utilized in order to improve intrusion detection, especially given the more sophisticated and evasive attack strategies. A study by Li *et al.* [11], developed DeepFed, a federated deep learning architecture for industrial cyber-physical systems, that used CNNs and Gated Recurrent Units (GRUs). DeepFed is particularly relevant as it highlights spatial and temporal features of network traffic, which can often be useful when looking for evasive threats, such as those that are able alter traffic patterns to evade detection. While focused on industrial systems, the deep learning architecture used in this study which is able to learn complex traffic behavioral is useful in adversarial resilience.

A study by Alarfaj and Khan. [1], used CNNs and LSTM networks to analyze network logs for anomaly detection. This work shows how deep architectures can learn complex traffic characteristics, which allows for a strong classification paradigm and could potentially be adapted to adversarial frameworks. In the case of attacker responses to introduce traffic that mimics normal or non-malicious behavior, the learning of "nuanced feature differences" learned by the deep networks can be leveraged. The issue of detecting malicious behavior through encrypted channels is taken on by Yang and Lim [18], which reassemble SSL records and find un-encrypted contents using deep learning. This method demonstrates the importance of meaningful feature selection from encrypted traffic, which tends to be a target of evasive behaviours. The ability for Deep Learning to be able to process sequential and encrypted data offers capabilities for resilient threat detection. In the case of anomaly detection, Song and Kim [16], propose a self-supervised approach that uses generator and detector models to create noised pseudo-normal data and to detect anomalies. This technique strengthens the model's robustness by exposing it to augmented data during training, which could provide some resistance to adversarial manipulations that aim to defeat detection. Self-supervised type techniques are important toward developing resilient systems that can respond to evolving threats. Studies by Azam *et al.* and Wong *et al.* [5, 17] discuss the move towards uncertainty quantification and Bayesian deep learning based models in the area of intrusion detection. Further Wong *et al.* [17] specifically develop models that provide predictions with uncertainty estimates, which is important for locating adversarial inputs which seek to exploit the predictions. It will augment frame works by using uncertainty measures to flag uncertain, or suspicious predictions that could be explored for resilience and

detection of adversarial inputs. Moreover, authors in study [9], achieved high accuracy detection of anomalies with deep learning using the MTA-KDD'19 dataset, showing that deep models can effectively identify other forms of traffic anomalies. Their work suggests that high detection capability can be achieved using well-trained deep architectures even in the context of the use of evasive tactics. A study by Mengara *et al.* [13], presented a hybrid deep learning model which utilized deep learning with some sort of method that can address the problems associated with class imbalance in the dataset and high-dimensional data which is typical in adversarial scenarios. Experimenting on datasets such as Botnet of Things-Internet of Things (Bot-IoT) and Canadian Institute for Cybersecurity Intrusion Detection System 2018 (CICIDS2018) dataset, showcased the effectiveness of hybrid models as a way to effectively increase the accuracy and robustness of detection in more complex settings involving a range of underlying processes. Together, these efforts demonstrate that the integration of advanced deep learning architectures include CNNs, LSTMs, transformers, and hybrid architectures, as well as self-supervision and uncertainty quantification, can markedly increase the robustness of intrusion detection systems to evasive cyber-attacks. Clearly designing and implementing such adversarially-resilient frameworks, is now an important avenue for future research in the areas of network security.

## 3. Methodology

The primary objective is to construct a robust system with high detection performance under normal as well as adversarially conditions. The proposed framework consists of the following major phases: dataset gathering and preprocessing, feature selection through RFE, hybrid deep learning model creation based on 1D-CNN and LSTM, adversarial training through the FGSM, and general model evaluation with an emphasis on robustness against adversarial attacks. The steps in the proposed approach is depicted in Figure 1.

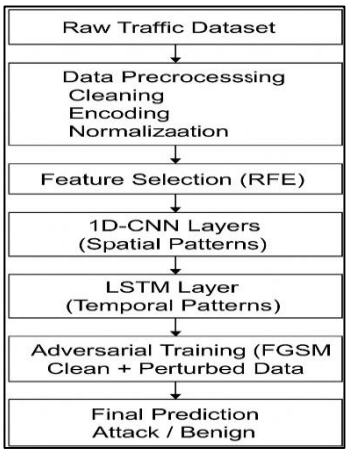


Figure. 1. Overview of the proposed approach.

### 3.1. Overview

The detection approach proposed in this study was developed to respond to the challanges of the correct detection of malicious behaviour on a network and do in an ordered manner while implemented defensively. The approach starts from preparation and preprocessing of raw network flow records. This includes the processes of identifying and removing unnecessary features, filling missing values, and encoding all categorical and numerical features. This step allows the input data to be in a normalized and processed manner, so that the deep learning model can analyze it successfully. After the preprocessing, the added functionality of RFE will be applied on the planned model architecture to determine the effective and informative features from the high-dimensional dataset. The basis of RFE is built around iteratively removing the least informative features while the model performance is selected and delivers a concise list of selected features. Not only does RFE lessen the computational burden on future training time, but it also helps improve the model performance through generalization, noise and redundant of data. A deep learning model was built using a hybridization of 1D-CNN and LSTM layers using preprocessed datasets.

CNN model is good for capturing local spatial relationships among features of the network, while the LSTM layers should identify temporal patterns such as sequential series, with both qualities significant for characterizing whether traffic is benign or malicious. Adversarial training is introduced to give the model some tolerance to input perturbations, leveraging the FGSM ability to create diverse input samples that simulate minor adversarial manipulation for evasion. Training with both clean and adversarial input samples allows the model to learn the character of inputs, when it has been deliberately manipulated, providing robustness in identifying them correctly. The trained model will be evaluated on clean and adversarial test datasets. Its performance is measured using accuracy, precision, recall, F1 score, and area under the curve Receiver Operating Characteristic-Area Under the Curve (ROC-AUC). This framework has the potential for achieving high detection rates under normal conditions and is embedded with reliability and robustness capabilities for utilization in potentially adversarial environments, thereby making it feasible for real-world applications related to cybersecurity.

### 3.2. Dataset Description

To evaluate the proposed intrusion detection model realistically and comprehensively, we will use the CICIDS2017 dataset [7], which is recognized as the standard dataset by the Canadian Institute for Cybersecurity. The CICIDS2017 dataset closely mimics the real traffic seen on enterprise networks over five days, and features a plethora of both benign and malicious flows. All traffic was generated by

CICFlowMeter, which analyzes raw packet data for flow-based features, thus allowing use in deep learning applications.

The CICIDS2017 dataset features approximately 3 million labeled records and has 14 various attack types, including Brute Force, PortScan, Distributed Denial-Of-Service (DDOS), Botnet, Infiltration, and web-based attacks. The traffic variations capture fluctuation throughout the day and during times when the traffic was transitioned to periods of inactivity, increasing the realism of the dataset. Each record provides over 80 features with multiple packets and bytes, protocol details, and many statistical measures. The volume and diversity found in the CICIDS2017 dataset provide an intricate foundation to the detection of attacks, especially long and subtle ones, through the use of deep learning. CICIDS2017 is an appropriate data set to evaluate the effectiveness in accuracy, generalizability, and adversarial robustness in a balanced and realistic way for deep learning-based intrusion detection model.

### 3.3. Data Preprocessing

Raw network traffic data usually contains deviations, noise, and redundant data that adversely affect the performance of Artificial Intelligence (AI) algorithms. Therefore, a structured preprocessing pipeline must be established to produce usable data. The first step in the preprocessing pipeline is to remove redundant and non-informative features such as flow ID, timestamp, source IP, destination IP, and string-formatted labels because they contain no information aiding learning and are a source of noise. The next step involves searching the dataset for missing or infinite values, as these are typical of real-world traffic logs. In each case of a missing or infinite value, we use median imputation to fix the problem, as this is a strong imputation technique that will not introduce skew to the data and helps preserve the general distribution. Categorical features such as protocol, state, and service are then converted into numerical representations via one-hot encoding. This process allows the neural network to interpret categorically represented variables (0and1) without establishing any false ordinal relationships for the variable. All continuous numbered features are scaled to a [0, 1] scale using min-max normalization, which helps establish standardized feature space, improves convergence during training, and eliminates failure associated with vanishing or exploding gradients. The preprocessing steps yield a clean, consistent dataset in numerical stable form that is applicable to training deep learning models.

### 3.4. Feature Selection Using Recursive Feature Elimination (RFE)

Although the CICIDS2017 dataset has multidimensional features, not all features contribute the same way to the classification. Therefore, to ease the model's complexity and lessen the threat of over fitting so RFE was used. RFE is an iterative feature ranking technique which selects the 30 most informative features according to its importance and removes the trailing important using a machine learning model for importance in this case we selected a random forest classifier because of its robustness and trustworthy feature importance capacity. Subsequently, all 80+features were included. In each round of the RFE, the least important features were removed based on their contribution within the Random Forest model to the model's predictive performance, until 30 features were remaining. Those 30 selected features were then used to train the deep learning model to ensure better performance in terms of both computational speed to convergence and classification accuracy.

### 3.5. Deep Learning Architecture

The core of the approach is a model that merges 1D-CNN for spatial feature learning and LSTM for temporal pattern extraction. This hybrid architecture lends itself well to the task of intrusions detection in networks, where flow-based features show local interactions and sequential tendencies.

#### 3.5.1 1D-CNN + LSTM Hybrid Model

The proposed modelling uses a hybrid deep learning architecture which blends one 1D-CNN with LSTM units to effectively capture both spatial and temporal dependencies in network traffic data. The model receives the feature vector corresponding to a single flow as input and is formatted as shape (1,30) where 30 is the number of features selected after preprocessing and feature selection. The first part of the model consists of a 1D-CNN layer that applies 64 convolutional filters all of size 3x1 and a stride of 1. This convolutional operation scans the input feature sequence as a moving kernel in order to learn local representations of features and patterns. Mathematically, the convolution operation for an input vector $\mathcal{X}$ and filter $\mathcal{W}$ is expressed as given in Equation (1).

$$\mathcal{Y}_i = \sum_{k=0}^{k-1} x_{i+j}\, \mathcal{W}_j \tag{1}$$

Where $k$ is the kernel size and $Y_i$ is the output at position $i$. The purpose of this layer is to detect correlations between adjacent features, which may indicate suspicious patterns in flow behavior. After convolution, a Rectified Linear Unit (ReLU) activation function is applied element-wise to introduce non-linearity into the model. The ReLU function is defined as given in Equation. (2).

$$ReLU(x) = max(0, x) \tag{2}$$

which ensures that only positive activations are propagated, helping the network to learn complex decision boundaries more effectively.

After ReLU activation, a maxpooling layer with pool size of 2 is applied to decrease the spatial dimensions of the feature map, this will help in the reduction of the number of parameters and the computational cost. In MaxPooling, the maximum values in its pooling window are selected in order to preserve the most important features. The pooled output is then taken as input to an LSTM layer with 100 hidden units. The LSTM networks have the ability to capture temporal sequences and long-term dependencies, which allows them to capture sequences of behaviours in a networked traffic such as the output of scanning attempts, or large bursts of coordinated attacks. The workings of an LSTM unit involve various gated processes that control the flow of information: Equation (3) represents the forget gate, Equation (4) the input gate, Equation (5) the cell state update, and Equation (6) the output gate.

***Algorithm 1:*** *Adversarially-robust deep learning for intrusion detection.*
***Require:*** *Network flow dataset D, perturbation magnitude ε*
***Ensure:*** *Trained and adversarially robust model M*

***Preprocessing Phase:***
*Remove redundant features from D*
*Impute missing values using median values*
*One-Hot encode categorical features*
*Normalize all numerical features using Min-Max scaling*
***Feature Selection:***
*Apply Recursive Feature Elimination (RFE) to D*
*Select top-k features to obtain reduced dataset D′*
***Adversarial Sample Generation (FGSM):***
***for*** *each (x, y) ∈ D′* ***do***
*Compute gradient* $g_x \leftarrow \nabla_x J(\theta, x, y)$
*Generate adversarial sample:* $x_p \square_v \leftarrow x + \varepsilon \cdot sign(g_x)$
***end for***
*Combine clean and adversarial samples:*
$D_{train} \leftarrow D' \cup D'_a \square_v$
***Model Training:***
*Train hybrid 1D-CNN + LSTM model M on* $D_{train}$
***Model Evaluation:***
*Evaluate M on clean test data* $D_n$
*Evaluate M on adversarial test data* $D_a \square_v$
***return*** *M*

$$f_t = \sigma(W_f . [h_{t-1}, x_t] + b_f) \quad (3)$$

$$i_t = \sigma(W_i . [h_{t-1}, x_t] + b_i), \tilde{C}_t = tanh(W_c . [h_{t-1}, x_t] b_c) \quad (4)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (5)$$

$$o_t = \sigma(W_i . [h_{t-1}, x_t] + b_i), h_t = o_t * tanh(C_t) \quad (6)$$

Here, $x_t$ is the input at time step $t$, $h_{t-1}$ is the hidden state from the previous step, $\sigma$ is the sigmoid function, and * denotes element-wise multiplication. These mechanisms allow the LSTM to retain or discard information across time steps, enabling the network to learn temporal attack signatures. The output of the LSTM layer is flattened and sent into a fully connected dense layer, which combines the learned features. Then the model predicts by outputting, through the Softmax activation function, probability scores over all attack classes. The *Softmax* function can be expressed as in Equation (7).

$$Softmax(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad (7)$$

where $z_i$ is the input to the $i$-$^{th}$ neuron in the output layer and $K$ is the number of classes. This function ensures that the output values lie between 0 and 1 and sum to 1, making them interpretable as class probabilities.

The architecture proposed captures the spatial feature extraction strengths of CNNs and LSTMs temporal modeling capabilities. The architecture is well suited for many types of network intrusion detection, whether short-term spikes in activity or long-term activity such as port scanning, denial-of-service, or multi-stage intrusion.

### 3.6. Adversarial Training with FGSM

Deep models are very powerful, but adversarial examples-inputs that are only slightly perturbed to avoid detection-pose one threat facing models. To combat this concern, we will apply adversarial training based on the FGSM during model training. FGSM perturbs input features in the direction of the model's loss gradient as given in Equation (8).

$$x_{adv} = x + \epsilon . sign\,(\nabla_x J(\theta, x, y)) \quad (8)$$

This study uses adversarial model used for creating and assessing FGSM based perturbations. A white-box adversary, possessing full information about the structure of the model and the gradients of the model allows the attacker to create a perturbation to their input and calculate the loss gradient for this modified input in order to generate an input perturbation that will misclassify the classifier's (model's) decision. This white-box adversary also has the capability to modify different network flow features during the testing phase, to create a limited input perturbation, such that the output remains characterized closely to the original input. The adversary is also limited to making input-level adjustments only; it cannot manipulate the model weights or the internal training data prior to generating the classifier's decision, nor can it change the class label that the classifier proposes (the results of the classification). The adversary's goal is to cause flows that were malicious to be identified as benign by the classifier while ensuring that the perturbation is minimized in size this is achieved through the use of FGSM. The FGSM attack serves as an excellent example of a strong first-order white-box attack, and it is a basis for evaluating and enhancing the adversarial robustness of algorithms.

## 4. Experimental Setup, Results, and Analysis

To validate the effectiveness and robustness of the proposed adversarial-resilient deep learning model for

intrusion detection, a complete series of exhaustive experiments were undertaken using the CICIDS2017 dataset. The model was implemented in Python utilizing the TensorFlow and Keras libraries. All experiments were done using a system that has an NVIDIA RTX 3080 GPU, 32 GB RAM, and Intel i9 processor running Ubuntu 22.04 LTS. This setup allowed for efficient training and evaluation of deep learning models on large datasets. Algorithm 1 details all the steps involved.

The processed dataset was split into 70% training data and 30% test data. An additional 10% of training data was held out for validation during training. Adversarial examples were generated with the FGSM using a perturbation size of $\varepsilon$=0.05 and inserted into the training and evaluation stages. Hyperparameters were tuned using grid search and the best hyperparameter values were batch size 256, learning rate 0.001, and value of 50 epochs for training. Dropout (with rate 0.3) and early stopping were implemented to mitigate overfitting. The performance of the model was evaluated using common metrics: Accuracy, Precision, Recall, F1-Score, and ROC-AUC. The metrics were computed on both clean (unperturbed) and adversarial perturbed test data to determine the robustness of the model. The results are shown in the Table 1.

Table 1. Experimental results.

| Metric | Clean test data | Adversarial test data |
|---|---|---|
| **Accuracy** | 93.1% | 86.3% |
| **Precision** | 92.5% | 85.4% |
| **Recall** | 91.8% | 83.6% |
| **F1-Score** | 91.7% | 84.5% |
| **ROC-AUC** | 0.96 | 0.89 |

The results indicate that the hybrid model achieves greater than 91% for all with an area under the Receiver Operating Characteristic (ROC) curve (AUC) of 0.96 on clean test data, which is indicative of very good discrimination. Most significantly, while also performing under an adversarial attack condition, the model exhibits good accuracy (86.3%) and a strong F1-score (84.5%). The ability of the model to maintain such performance under an adversarial condition is representative of the FGSM adversarial training and its impact on ultimately learning more generalizable and robust decision boundaries. The impact of using RFE in the feature selection process to improve all models performance. When the input feature space was reduced to the 30 most relevant features RFE helped remove noise and highly correlated/redundant features and ultimately led to an improved model performance via more rapid convergence during training as well as improved generalization performance to unseen data. Dropout regularization also lessened the risk of any overfitting as evidenced by the model's performance on validation and test data, which were very consistent.

To further compare the results, the hybrid model was also compared against widely used appropriately baseline classifiers, SVM, Random Forest (RF) and vanilla Multi-Layer Perceptron (MLP) using the same preprocessing, features selection process and adversarial attack testing process as the hybrid model. The results can be seen as shown with model comparisons based on F1-score in Table 2.

Table 2. Comparison with different models.

| Model | F1-score (clean) | F1-score (adversarial) |
|---|---|---|
| **SVM** | 82.1% | 61.4% |
| **Random forest** | 88.3% | 70.2% |
| **MLP (vanilla)** | 86.4% | 72.6% |
| **Proposed model** | **91.7%** | **84.5%** |

The proposed hybrid model performed significantly better than all baseline models, especially in the adversarial setting. While baseline models (SVM and RF) sustained significant performance drops, SVM down to 61.4% F1-score and (RF) down to 70.2% F1-score, the hybrid model only reduced performance by 7.2%, indicating its ability to detect evasive attack behavior is better and points to the advantages of adversarial training for enhancing model robustness.

Overall, the experimental outcomes establish that the proposed 1D-CNN + LSTM hybrid model, with feature optimization and adversarial training provides high accuracy and strong robustness in network intrusion detection. It provides a strong level of defence against not only traditional attack types but also against adversarially-crafted inputs produced to evade intrusion detection systems. These results suggest that the model is capable of being utilized in a real-world security context where robustness and reliability are considered crucial.

## 5. Conclusions

This study demonstrates a robust and scalable deep learning framework to identify evasive cyber threats in network traffic with an emphasis on resistance to adversarial attacks. By effectively combining the spatial feature extraction capabilities of the 1D-CNN method with the sequential learning method (LSTM trained), the model is able to identify the inherent structure and temporal behavior of network flows. In addition, RFE improves both computational efficiency and generalization by removing redundant or noisy features. Finally, adversarial training with the FGSM enhanced robustness by training it to identify adversarial perturbations by attackers. The experimental evaluation using CICIDS2017 demonstrated that the hybrid model had not only great performance in normal conditions-scoring greater than 91% across all standard classification metrics-but also remained strong through adversarial conditions, suffering only a reduction in F1-score. Through this experiment the hybrid model outperformed both in clean and adversarial testing conditions against traditional models like SVM, Random

Forest, and MLP, confirming it's superiority for defensive anti-cyber threat applications, and furthermore confirming its efficacy as a practical tool in real-world situations. The results of this study fill a gap in cybersecurity in respect to adversarial robustness in deep learning-based intrusion detection systems. This work opens the door to future deployment of intelligent, adaptive, and secure IDS in high value environments by showing that adversarial training, feature optimization and hybrid neural networks provide the necessary defences against complex, evasive cyber threats.

## Acknowledgment

The authors would like to acknowledge the Deanship of Graduate Studies and Scientific Research, Taif University for funding this work.

## References

[1] Alarfaj F. and Khan N., "Enhancing the Performance of SQL Injection Attack Detection Through Probabilistic Neural Networks," *Applied Sciences*, vol. 13, no. 7, pp. 1-11, 2023. DOI: 10.3390/app13074365

[2] Alqarni A., Alsharif N., Khan N., Georgieva L, and et al, "MNN-XSS: Modular Neural Network Based Approach for XSS Attack Detection," *Computers, Materials and Continua*, vol. 70, no. 2, pp. 4075-4085, 2022. DOI: 10.32604/cmc.2022.020389

[3] Alshehri A., Khan N., Alowayr A., and Alghamdi M., "Cyberattack Detection Framework Using Machine Learning and User Behavior Analytics," *Computer Systems Science and Engineering*, vol. 44, no. 2, pp. 1679-1689, 2023. DOI: 10.32604/csse.2023.026526

[4] Al-Shurbaji T., Anbar M., Manickam S., Hasbullah I., and et al., "Deep Learning-Based Intrusion Detection System for Detecting IoT Botnet Attacks: A Review," *IEEE Access*, vol. 13, pp. 11792-11822, 2025. DOI: 10.1109/ACCESS.2025.3526711

[5] Azam Z., Islam M., and Huda M., "Comparative Analysis of Intrusion Detection Systems and Machine Learning-Based Model Analysis Through Decision Tree," *IEEE Access*, vol. 11, pp. 80348-80391, 2023. DOI: 10.1109/ACCESS.2023.3296444

[6] Bansal K. and Singhrova A., "Enhanced IIoT security: A Hybrid Intrusion Detection Framework Using Signature-Based and Deep Learning Approaches," *The International Arab Journal of Information Technology*, vol. 23, no. 2, pp. 233-245, 2026. DOI: 10.34028/iajit/23/2/5

[7] Canadian Institute for Cybersecurity, Intrusion Detection Evaluation Dataset (CIC-IDS2017), https://www.unb.ca/cic/datasets/ids-2017.html, Last Visited, 2025.

[8] El-Amir S., "Comprehensive Cybersecurity Review: Modern Threats and Innovative Defense Approaches," *International Journal of Computers and Informatics (Zagazig University)*, vol. 1, pp. 30-37, 2023. https://www.ijci.zu.edu.eg/index.php/ijci/article/view/77

[9] Khalaf L., Alhamadani B., Ismael O., Radhi A., and et al., "Deep Learning-Based Anomaly Detection in Network Traffic for Cyber Threat Identification," *in Proceeding of the 24th Cognitive Models and Artificial Intelligence Conference*., Istanbul, pp. 303-309, 2024, DOI: 10.1145/3660853.3660932

[10] Khan N., Abdullah J., and Khan A., "Defending Malicious Script Attacks Using Machine Learning Classifiers," *Wireless Communications and Mobile Computing*, vol. 2017, pp. 1-9, 2017. DOI: 10.1155/2017/5360472

[11] Li B., Wu Y., Song J., Lu R., and et al., "DeepFed: Federated Deep Learning for Intrusion Detection in Industrial Cyber-Physical Systems," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 8, pp. 5615–5624, 2021. DOI: 10.1109/TII.2020.3023430

[12] Liu Q., Hagenmeyer V., and Keller H., "A Review of Rule Learning-Based Intrusion Detection Systems and their Prospects in Smart Grids," *IEEE Access*, vol. 9, pp. 57542-57564, 2021. DOI: 10.1109/ACCESS.2021.3071263

[13] Mengara A., Yoo Y., and Leung V., "IoTSecUT: Uncertainty-Based Hybrid Deep Learning Approach for Superior IoT Security Amidst Evolving Cyber Threats," *IEEE Internet of Things Journal*, vol. 11, no. 16, pp. 27715-27731, 2024. https://ieeexplore.ieee.org/document/10538199

[14] Sagar R., Jhaveri R., and Borrego C., "Applications in Security and Evasions in Machine Learning: A Survey," *Electronics*, vol. 9, no. 1, pp. 1-42, 2020. DOI: 10.3390/electronics9010097

[15] Sanikommu V., Mummaneni S., Jacob N., Emmanuel K., and Variam R., "Deep Learning-Based Control System for Context-Aware Surveillance Using Skeleton Sequences from IP and Drone Camera Video," *The International Arab Journal of Information Technology*, vol. 22, no. 5, pp. 919-929, 2025. DOI: 10.34028/iajit/22/5/6

[16] Song H. and Kim H., "Self-Supervised Anomaly Detection for In-Vehicle Network Using Noised Pseudo Normal Data," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 2, pp. 1098-1108, 2021. DOI: 10.1109/TVT.2021.3051026

[17] Wong J., A. M. Berenbeim, Bierbrauer D., and Bastian N., "Uncertainty-Quantified, Robust Deep Learning for Network Intrusion Detection," *in*

*Proceeding of the. Winter Simulation Conference. (WSC)*, Texas, pp. 2470-2481, 2023. DOI: 10.1109/WSC60868.2023.10407559

[18] Yang J. and Lim H., "Deep Learning Approach for Detecting Malicious Activities over Encrypted Secure Channels," *IEEE Access*, vol. 9, pp. 39229-39244, 2021. DOI: 10.1109/ACCESS.2021.3064561

**Nasser Alsharif** is an Assistant Professor of Artificial Intelligence and Computer Science at Taif University, Saudi Arabia. He holds a Ph.D. in Trust Evaluation and Management in eCommerce Systems using Artificial Intelligence and Emerging Technologies from the University of the West of Scotland, UK, and a Master's degree in Business Information Technology from Edinburgh Napier University, UK. With over a decade of experience in higher education and academic leadership, his research interests include Artificial Intelligence Applications, Electronic Systems Security, Embedded Technologies, Quality Assurance, and Educational Innovation. Dr. Alsharif has also held several academic and administrative leadership roles, contributing to curriculum development, accreditation, strategic planning, and institutional advancement.