



An Explainable Deep Learning Approach for Distinguishing Cyber Attacks and Sensor Faults in Critical IoT Systems

Nabeel I. Zanoon¹, Abdullah Odeh Al-Zaghameem², Khalid Alkharabsheh³

¹Department of Applied Sciences, Al-Balqa Applied University, Aqaba, Jordan

²Department of Information Technology, Tafila Technical University, Tafila, Jordan

³Department of Software Engineering, Al-Balqa Applied University, As-Salt, Jordan

E-mails: dr.nabeel@bau.edu.jo aoz@ttu.edu.jo khalidkh@bau.edu.jo

Abstract: The increasing rate of cyber-attacks in critical Internet of Things (IoT) systems requires robust and interpretable detection systems. In this paper, an interpretable deep learning approach has been introduced to detect various types of cyber-attacks and distinguish them from sensor faults. The proposed model uses Convolutional Neural Networks-Bidirectional Long Short-Term Memory (CNN-BiLSTM) to extract features. The attention mechanism has also been employed to improve the interpretability of the model. In addition, the SHAP approach has been used to interpret the model's predictions by measuring the contribution of features. The performance of the model has been evaluated using the TON_IoT dataset with an accuracy of 97.3% for multi-class classification and 94.25% for binary classification. The results show that the proposed approach offers high detection accuracy and interpretability of the results.

Keywords: Internet of Things (IoT) System, SHAP, Convolutional Neural Networks (CNN), Bidirectional Long Short-Term Memory (BiLSTM), EXplainable Artificial Intelligence (XAI), Cyberattacks.

1. Introduction

The Internet of Things (IoT) is an advanced technology that has made its way into a variety of industries. Millions of smart devices communicate in real time to collect and process data. In these industries, devices use IoT sensors to predict equipment failures or report problems [1]. The increased device count could lead to an overall increase in risk and vulnerability related to these IoT Systems. Cybersecurity aims to protect the entire ecosystem of connected devices from cyberattacks, as well as to

deter anyone from accessing any device without authorization [2]. An IoT system is made up of many devices that are constantly connected to the internet, and these devices represent potential vulnerabilities that could be exploited by malicious actors. Attack vectors on IoT devices include, but are not limited to, malware, ransomware, Denial of Service (DoS), Man-in-the-Middle (MitM), and phishing. These and other types of attacks may allow an attacker to gain access to sensitive data and create severe security risks [3]. New technologies like sensors, smartphones, wireless communications, cloud computing, etc., represent an Internet of Things (IoT) system where devices send messages to each other or interact with each other. Due to this continued growth and rapid development of the multiple pieces of systems, new vulnerabilities for attacks to occur on or through these devices (or even within the system itself) could emerge. Unfortunately, traditional methods for preventing and protecting systems from hacker attacks or other forms of digital crime are ineffective against modern digital crime because of the nature of today's IoT systems which provide additional access points for criminals to take advantage of [4].

The ability to interpret information provided by the sensors about the physical environment will determine how safe and reliable they operate within their environment. The problem with interpreting sensor information is that sensors can be damaged or compromised, which can create sensor faults or failures, which can produce incorrect information about the current state of the environment. Determining whether there are issues with sensors is an important step in protecting IoT systems from hacker attacks and ensuring a safer system [5]. Detecting attacks is challenging because of their heterogeneity. The sheer volume of data, along with the dynamic nature of network traffic in IoT systems, complicates the detection process. To distinguish between harmless abnormalities and malicious behavior, intelligent models must be capable of comprehending and analyzing context. Furthermore, attackers are always creating new evasion techniques, making it difficult to detect assaults using typical network monitoring technologies. To address these challenges, extremely efficient and accurate intrusion detection systems that work in real time and can analyze patterns are required. All these challenges demand powerful deep learning solutions that can learn, adapt, and respond to emerging threats in real time, thereby enhancing IoT security [6].

In specific, deep learning models combine different Convolutional Neural Networks (CNN) and Recurrent Neural Network (RNN) techniques to support the detection of complex patterns and features in network traffic data through Long Short-Term Memory (LSTM) networks and Bidirectional Long Short-Term Memory (BiLSTM) networks. Using deep learning models, detection systems can

accurately identify anomalies in the data, resulting in identifying security threats. Deep learning-based IDS have demonstrated high accuracy and resilience to current threat vectors [7]. Furthermore, it allows for the detection of complex and emerging cyber threats in dynamic IoT environments. Deep learning-based IDS facilitates better comprehension of network activity and therefore allows security personnel to respond more effectively to threats. The integration of this technology enhances the efficiency of threat detection and strengthens cybersecurity infrastructure [8]. On the other hand, deep learning models act as black boxes by providing no clear explanation of their decision-making processes for IoT within many critical domains where decisions are required to be made with high precision. The absence of an explanation results in a lack of trust and transparency in the system [9]. EXplainable Artificial Intelligence (XAI) is a subfield of AI that facilitates the generation of highly accurate predictions combined with clear explanations of how the predictions were formed, thus providing improved trust and transparency to the end-user [10]. Both LIME [11] and SHAP [12] are frequently used interpretation techniques. Despite their methodologies, LIME is a local interpretation technique, while SHAP provides global as well as local interpretation techniques through its dual interpretation approach [13]. The effectiveness of the IDS is increased in terms of the identification of vital features using a technique that detects complex patterns of an attack on IoT network data without considering more noise and false alarms [14]. In this work, we exploit the power of deep learning techniques in terms of cyberattacks. To this end, we aimed to:

- Developing an interpretable hybrid framework based on CNN-BiLSTM and an Attention mechanism to distinguish cyberattacks from sensor failures in IoT environment.
- Enhancing spatiotemporal features extraction by coupling CNN to extract spatial features and BiLSTM to capture temporal dependence between features in the intricate IoT data.
- Strengthening interpretability and transparency of model decision by combining an Attention mechanism with SHAP to attain global and local explanation of decisions.
- Evaluate the efficiency and stability of the model by training the proposed model on a TON_IoT dataset specialized in the IoT environment and using comprehensive evaluation metrics such as MCC and Cohen's Kappa to assess classification accuracy, robustness, and reduction of false alarms.

2. Related work

Several previous studies on cybersecurity have focused on the application of explainable deep learning models for detecting cyberattacks on critical IoT systems, with many studies integrating Convolutional Neural Networks (CNNs) with LSTM and BiLSTM to exploit spatial and temporal features.

Sadhvani et. al. [15] present a deep learning-based hybrid intrusion detection system that combines BiLSTM and CNN networks to detect cyberattacks in IoT networks using UNSW-NB15 dataset. To improve performance, normalization and dropout optimization techniques were used. The proposed hybrid model outperformed expectations in terms of accuracy, detection rate, and false alarm reduction. The study recommends applying the proposed model to real data and various complex datasets and further developing for multi-category classification.

Similarly, the study in [16] introduced a hybrid CNN-BiLSTM transformer model that was presented to detect attacks in IoT environments. To solve data imbalance, the Borderline-SMOTE technique was used, and to increase model accuracy, an integrated feature selection mechanism using XGBoost, Chi-square, and Mutual Information. Experiments were carried out on the CIC-IDS2017 and BoT-IoT datasets, resulting in high performance. However, it was highlighted that the approach is computationally expensive and relies on non-real-time data. The paper proposed developing real-time IDS and implementing them in fields such as Industrial IoT (IIoT) and Vehicular Ad-hoc Networks (V2X).

In a more specialized study [17] focused on detecting DDoS attacks in IoT networks by proposing a hybrid model (CNN and LSTM), which used feature reduction based on the correlation coefficient; the model was evaluated on CICIoT2023 and TON_IoT datasets. In contrast to prior studies, the study conducted in [18] employs a different strategy, combining the attention mechanism with a hybrid model. Researchers propose an intrusion detection model that combines LSTM and ResNet to improve feature extraction and generalizability in IoT. The attention mechanism was integrated to focus on important features and reduce the impact of irrelevant elements. The model was evaluated on the TON_IoT and UNSW-NB15 datasets.

Saiyed, Al-anbagi and Hossain [19] proposed an explainable deep learning system for Internet of Energy (IoE) systems, called E-SHIELD, that aims to detect various cyberattacks in these systems. Their solution was built using a hybrid CNN and LSTM architecture and evaluated on two datasets, WUSTLIoT-2021 and CICIoT-2023. For monitoring purposes, the proposed solution provided a

dashboard that allows users to analyze predictions, importance of features, and real-time risk levels.

The research conducted by Shtayat et. al. [20] introduced a novel deep learning model to detect cyber-attacks on Electric Vehicle Charging Systems (EVCS). The approach has evaluated four deep learning models (namely, DNN, CNN, LSTM, and a hybrid CNN-LSTM model) on the CIC EV Charger Attack Dataset 2024 (CICEVSE2024). The results of that approach showed that the hybrid CNN-LSTM outperformed the other three models and attained an accuracy of 97.52%, precision of 97.55%, recall of 97.52%, and an F1-score of 97.53%.

However, the proposed approach of this research differs from all previous ones due to the integration of detection accuracy and interpretability in a single model. The proposed model is useful in extracting key information from IoT data using a CNN-BiLSTM architecture. At the same time, it is supposed to help to point out the most important pieces of information using an attention mechanism and interpretability techniques like SHAP. The study is significant because it differentiates cyberattacks from sensor failures using real and complex datasets (TON-IoT [21]).

3. Deep learning, IOT, and XAI

Deep learning models have been good at lots of different things. When you add them to IoT, they help beef up the security level. Deep learning algorithms have the capability to automatically identify features from raw sensor data. Such a feature is essential in the case of IoT applications that are handling unstructured, noisy, or complex network data, as well as critical applications like live monitoring or making predictions for maintenance [22, 23]. Furthermore, the collaboration between deep learning models and IoT is making devices smarter and capable of performing complex tasks. Such tasks include analyzing sensor data from IoT devices, allowing these devices to make decisions and predictions using data while improving efficient and accurate data processing [24].

In networks, deep learning models use multiple networks, CNNs, RNNs, LSTMs, and BiLSTMs to help accomplish the task of network intrusion detection. CNNs are effective at extracting local and spatial features. However, they have limited capability in learning temporal patterns. RNNs are designed to capture sequential data and temporal relationships. However, they suffer from vanishing and exploding gradient problems. LSTM networks address these limitations through a gate mechanism that enables learning from long-term dependencies; however, LSTM can only learn sequence features from one direction. In contrast, a

Bidirectional Long Short-Term Memory (BiLSTM) network can learn in two forward and backward directions, as it is able to extract only temporal features and ignore spatial ones. As a result, current mainstream research has integrated the BiLSTM network and CNNs to capture both spatial and temporal features effectively [25, 26]. Despite their advantages, LSTM networks are prone to overfitting, especially when there are a large number of parameters compared to the data volume. In addition, they are relatively slower in classification tasks due to the reliance of their layers on sequential operations, which require the execution of each of its steps based on the results of previous steps [27].

The BiLSTM network can address these problems by leveraging both past and future context in the data sequence [28]. The BiLSTM model outperforms LSTM in IoT security applications due to its bidirectional sequential data processing. This model analyses the entire temporal context of network traffic, which helps to detect complex patterns of multi-stage attacks and correlates early indicators with future results [29]. In support of these previous studies, which compared the LSTM and BiLSTM models on multiple datasets, the results showed that the BiLSTM model was more accurate and better distinguished attacks from normal traffic [30-34].

With the growth of IoT technology across many industries, there has been a need for XAI, which promotes trust in AI-based decisions in IoT-based systems [35]. Lots of deep learning systems are ‘black boxes’ in that their decision-making processes are opaque and difficult to interpret, creating challenges for developers who must understand them, as well as the potential for hidden biases and/or inaccuracies. Therefore, XAI solutions offer non-model-specific explanations of models (as an alternative to actual use) in IoT systems for the purpose of improving understanding of the model’s decisions; hence, providing improved transparency and trust [36].

To provide transparency and reliability for deep learning models used in IoT applications, post hoc model-agnostic XAI approaches are integrated with these models. LIME and SHAP are two of the most established post hoc XAI techniques that provide insight into the predictions of complex models by analyzing the influence of features on their associated predictions [37]. SHAP is a model-agnostic method for explaining models’ predictions that uses cooperative game theory to produce Shapley values for how much each feature contributes to an individual’s prediction and how features contribute globally to a model [38].

4. Research methodology

4.1. Experimental setup and dataset description

In the present work, the TON-IoT network dataset has been used. This realistic dataset was generated specifically for the purpose of emulating IoT and IIoT environments regarding critical systems. It contains a range of data, including network traffic, Windows and Ubuntu operating system logs, and measurements from IoT and IIoT sensors. Therefore, it can be used to study normal behavior, cyberattacks, and sensor failures (Table 1). The Train_Test_Network.csv file, containing 461,043 records of network traffic data, was used in all experiments.

Table 1. Class distribution in the TON-IoT dataset

Class	Number of samples
Normal	300,000
Backdoor	20,000
DDoS	20,000
DoS	20,000
Injection	20,000
Password	20,000
Ransomware	20,000
Scanning	20,000
XSS	20,000
MITM	1043

The TON-IoT dataset has a class imbalance, where the normal behavior class has (300,000) records, and each of the attack classes has only (20,000). Among these attack classes, the Man-in-the-Middle (MitM) class has the least number of records available (1043), which is an example of how class imbalances are common in IoT security datasets. The class imbalance makes multi-classification of classes more complex for the proposed model. To mitigate the effect of class imbalance, error functions were combined with class weights to reduce the bias the model may show based on the size of the class being represented. The effects of data imbalance on the performance of the model will be discussed in later sections.

4.2. Data preprocessing and partitioning

Before model training, the dataset underwent some preprocessing operations for data duplication removal, missing data correction, feature deletion, and non-numerical feature conversions to create appropriate numerical data representation. The numerical data was then normalized using min-max scaling for learning stability and consistency.

Data segmentation is a method that divides data into learning and evaluation subsets to produce a reliable and generalizable machine learning model. This allowed 70% of the dataset to be divided into training and 30% into testing.

Therefore, 322,730 records will be used in the training process and 138,313 records in the testing process.

The training data was utilized to train the model and provide parameters, and the testing data was used solely to assess final performance in terms of binary and multi-class classification, without impacting model training. The data segmentation and training settings are shown in Table 2.

Table 2. Dataset partitioning and CNN-BiLSTM-attention training configurations

Item	Description / Value
Dataset	TON-IoT Network (Train_Test_Network.csv)
Total samples	461,043
Training set	70% (322,730 samples)
Testing set	30% (138,313 samples)
Classification type	Binary and multi-class
Number of classes	2 (Binary), 10 (multi-class)
Number of features	45
Normalization	Min-Max Scaling
Model	CNN-BiLSTM-Attention
Optimizer	Adam
Learning rate	0.001
Batch size	64
Epochs	50
Loss function (Binary)	Binary Cross-Entropy
Loss function (multi-class)	Categorical Cross-Entropy
Early stopping	Enabled

The framework developed in the study, as illustrated in Fig. 1, provides an explainable hybrid deep learning model that incorporates efficient spatial and temporal feature extraction mechanisms in an explainable artificial intelligence framework. In developing the framework, the methodology aims to provide a comprehensive analysis of cyberattacks and sensor faults in critical IoT environments. For example, the framework uses CNNs to effectively learn discriminative spatial features from heterogeneous data streams. In addition, BiLSTM networks effectively capture short-term and long-term temporal dependencies from sequential IoT data, enabling the model to effectively capture dynamic behavior.

Furthermore, to provide more meaningful temporal insights into the emergence of anomalous behavior, an attention mechanism is integrated to effectively capture the most significant temporal segments contributing to the model predictions. Finally, SHAP is integrated to effectively quantify the features contributing to each classification outcome, providing transparent and human-interpretable explanations for each outcome. By effectively integrating deep learning and XAI mechanisms, the framework provides enhanced detection capabilities as applied to TON_IoT dataset while ensuring trustworthy decision

support to distinguish between cyberattacks and normal sensor faults in critical IoT environments.

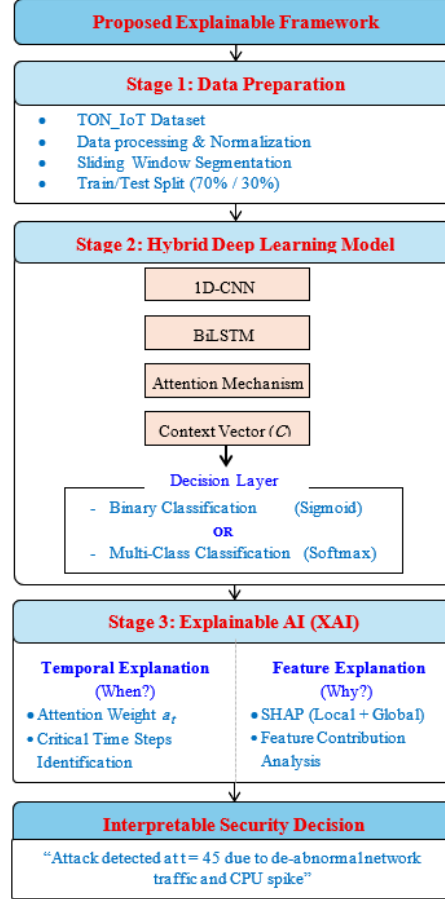


Fig. 1. The proposed model architecture

The proposed design for the model processes temporal, multidimensional IoT data from the TON_IoT dataset. The input into the system is a multivariate time series, as represented by the equation

$$(1) \quad X = \{x_1, x_2, \dots, x_T\}, x_t \in \mathbb{R}^d,$$

where T is the length of the time window used to analyze the data, while d is the number of features retrieved from the readings of various sensors at each point of time.

The multidimensional temporal representation allows the proposed hybrid model to capture spatial correlations between distinct attributes as well as temporal dependencies over the data sequence, where it serves as the basis in sequential processing over model stages.

The proposed model architecture is structured into multiple stages. In the first stage, CNN layers extract spatial features from input data by applying one-dimensional convolution operations on the time dimension according to the equation

$$(2) \quad F_k = \phi(W_k * X + b_k),$$

where: W_k is the k -th convolution filter; b_k is the bias vector; $\phi(\cdot)$ is a nonlinear activation function, specifically, the ReLU function, which contributes to improving the quality of the spatial representations extracted from the CNN layers. The extracted spatial representations are considered high-level and more stable and are used as input to the temporal analysis phase in the BiLSTM model; thereby enhancing its ability to learn long- and short-term time dependencies.

The BiLSTM network will capture the long- and short-term temporal dependencies in both the forward and backward directions, where the hidden state is determined at time t as given in the following equations:

$$(3) \quad \vec{\mathbf{h}}_t = \text{LSTM}_f(\mathbf{F}_t, \vec{\mathbf{h}}_{t-1}),$$

$$(4) \quad \tilde{\mathbf{h}}_t = \text{LSTM}_b(\mathbf{F}_t, \tilde{\mathbf{h}}_{t+1}).$$

The final time representation is produced by combining the outputs from both directions, as shown in the equation

$$(5) \quad \mathbf{h}_t = [\vec{\mathbf{h}}_t; \tilde{\mathbf{h}}_t].$$

This bidirectional representation contributes to the detection of sudden temporal patterns associated with cyberattacks, as well as gradual changes that indicate sensor failures. The model analyzes the time series from start to end (forward) and from end to start (backward). It does not only consider past events at the current step; this plays an important role in IoT attacks, as an attack may begin with one pattern and end with another pattern, so studying the sequences bidirectionally enables distinguishing between a sudden malfunction and an organized attack.

In the third stage, integrating the attention mechanism aims to improve the model's focus on the most critical time steps and to calculate a relative weight for each temporal step e_t based on the time representation result from the BiLSTM layers, as defined in the following equations:

$$(6) \quad e_t = \mathbf{v}^T \tanh(\mathbf{W}_a \mathbf{h}_t + \mathbf{b}_a),$$

$$(7) \quad \alpha_t = \frac{\exp(e_t)}{\sum_{i=1}^T \exp(e_i)}.$$

The unnormalized significance score for time step t denoted by e_t , is calculated using the BiLSTM output, where T represents the duration of the time series. The exponential function, $\exp(\cdot)$, ensures that all weights are positive.

The SoftMax function ensures that all attention weights are positive and normalized so that their sum equals one. Each α_t weight reflects the relative importance of the corresponding time step in the model's final decision. This normalization enhances model ability to identify time intervals associated with anomalous behavior, whether caused by organized cyberattacks or physical sensor faults, while reducing the impact of irrelevant time intervals. The SoftMax function converts unnormalized values into relative attention weights (α_t) and uses them to compute and constructs the final contextual vector $\mathbf{C}$, as shown in the equation

$$(8) \quad \mathbf{C} = \sum_{t=1}^T \alpha_t \mathbf{h}_t.$$

The $\mathbf{C}$ time representation vector will be fed into the multi-class classification step, where it will facilitate temporal interpretation by analyzing attentional weights to identify when anomalous behavior occurred in the system. At this stage, multi-class classification and fault differentiation are performed. The weighted contextual vector $\mathbf{C}$ is received and transferred to a fully connected layer that uses a SoftMax function for multi-class classification. The purpose of this classification is to distinguish between normal behavior, different types of cyberattacks, and physical sensor failures.

Precise decisions that can always be trusted are necessary to provide support for high-risk environments; however, classifying items into multiple classes and recognizing faults is not enough to provide effective decision support in multi-class classifications/recognition of faults. This research elucidates a new design for transforming a mechanism ("black box") into a highly efficient "transparent" system utilizing XAI. The approach is considered in its design, combining two levels of complementary interpretations to provide the needed transparency and reliability levels of decision-making for the created model.

Additionally, the innovative hybrid attention based on temporal interpretation provides an innovative mechanism for temporal interpretation outside of and in addition to traditional hybrid attention models. Each "step" in the input sequence during stage three generates a relative weight (e_t) assigned to each date/time that will determine the overall impact of that date/time on the overall output decision of the created model. By identifying critical time intervals before the occurrence of either an attack or malfunction, the created model can differentiate between cyberattacks that occur in structured, sequentially based timeframes and sensor faults that occur in unpredictable/abruptly/gradually and/or have no association with the operation of the network. By providing a temporal understanding of when the sensor fault or cyberattack has occurred, this capability will enhance the detection, monitoring, and effective (near real-time) response to faults within a safety-critical system.

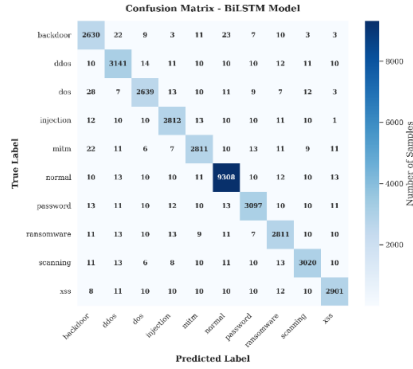
The proposed model uses SHAP for both local and global interpretation of decisions made by the model. Local interpretations determine how all features for an individual prediction impact a feature by examining the influence of the feature on this individual prediction. Global interpretations determine how all features impact predictions across the entire dataset by looking at the overall contribution of each feature across all feature/label pairs. This enables users to distinguish between network features that could indicate a cyberattack and sensor features that could indicate a normal physical fault. Additionally, by identifying the impact of features on the overall classification decision, SHAP provides a way to articulate why a model produces a particular output for a given input.

5. Results and discussion

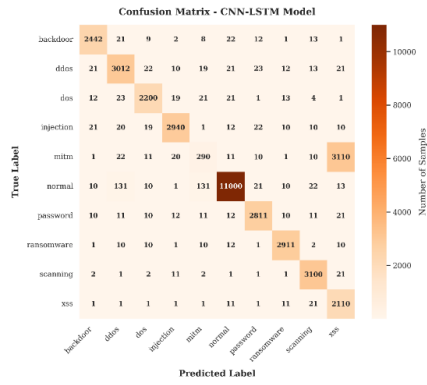
5.1. Comparison of deep learning models for multi-class classification

To evaluate the effectiveness of the proposed model, its performance was compared with those of the BiLSTM and CNN-LSTM models using multi-class confusion matrices, as shown in Fig. 2. The BiLSTM model, shown in Fig. 2a, demonstrated good capability in learning bidirectional temporal dependencies. However, the lack of a dedicated feature extraction mechanism limited its ability to distinguish between similar attack patterns, resulting in a significant number of misclassifications away from the main diagonal.

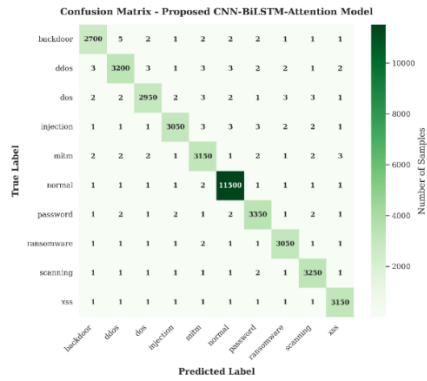
The CNN-LSTM model, shown in Fig. 2b, benefits from the combination of CNN-based spatial feature extraction and LSTM-based temporal dependency modeling, resulting in better overall classification performance. However, there was significant confusion between the MITM and XSS classes. Only 290 MITM samples were accurately classified, whereas 3110 MITM samples were misclassified as XSS. This result indicates that the model struggled to learn sufficiently discriminative representations for MITM attacks, leading to substantial class overlap between these two attack categories. Additionally, 131 samples of normal traffic were misclassified as DDoS attacks, leading to higher false positive rates. Fig. 2c shows that the CNN-BiLSTM-Attention model achieves the highest-class differentiation, with most values centered on the main diagonal of the confusion matrix and a significant drop in misclassification values beyond the diagonal. It also achieved 11,500 correct classifications for the Normal class and 3150 correct classifications for the MITM class, with a significant reduction in confusion associated with the XSS class compared to other models.



(a) Confusion matrix of the baseline BiLSTM model



(b) Confusion matrix of the hybrid CNN-LSTM model



(c) Confusion matrix of the proposed CNN-BiLSTM-Attention model

Fig. 2. Performance comparison via multi-class confusion matrices for the evaluated network intrusion detection models: the baseline BiLSTM model (a); the hybrid CNN-LSTM model (b); and the proposed attention-driven CNN-BiLSTM-Attention model (c)

Overall, these results confirm that the integration of CNN-based feature extraction, BiLSTM-based bidirectional long short-term memory, and the attention mechanism improved the model's ability to distinguish between classes and overall

classification performance, despite the relative imbalance in class distribution shown in Table 1.

5.2. Multi-class performance evaluation

A quantitative evaluation was conducted to assess performance against a set of standard classification performance metrics shown in Table 3. The results demonstrate that the CNN-BiLSTM-Attention model achieved a greater performance than expected compared to the CNN-LSTM and BiLSTM models, with 97.3% classification accuracy vs. 93.3% and 89.7%, respectively. It gives a strong indication that this model has an outstanding ability to accurately distinguish between multiple classes in a multi-class classification scenario.

Table 3. Classification performance metrics of the proposed models in the multi-class

Metric	BiLSTM model (a)	CNN-LSTM model (b)	CNN-BiLSTM-Attention (c)
Accuracy (Overall)	89.7%	93.3%	97.3%
Precision (Macro Avg)	88.6%	92.4%	96.7%
Recall / Detection rate (Macro Avg)	90.3%	94.0%	97.5%
F1-score (Macro Avg)	89.2%	93.0%	97.0%
FAR (%) (False Alarm Rate)	1.25%	0.82%	0.34%

Through analysis, the proposed model produced the highest average macro precision (96.7%), recall/detection rate (97.5%), and F1-score (97.0%), indicating a high degree of balanced performance and efficiency across all categories. Additionally, there was also a substantial decrease in false negatives and positives. Moreover, the model produced the lowest FAR (0.34%), which increases the reliability of operations in critical environments, validating that this model can be used in effective cyberattack detection systems in mission-critical IoT systems. Therefore, these findings provide evidence of the ability of this model to provide high levels of accuracy, a balance between recall and precision, and low false alarms, making it suitable for use in high error environments.

5.3. Multi-class statistical metrics comparison

Several advanced statistical metrics were assessed to enhance the reliability of the evaluation as reflected in Table 4, for instance, Matthews Correlation Coefficient (MCC) and Cohen's Kappa. Between the models, the recommended CNN-BiLSTM-Attention model outperformed the comparators in both the Matthews Correlation Coefficient (MCC = 0.969) and the Cohen's Kappa (0.970) values, supporting that the recommender did indeed demonstrate the best overall accuracy.

Table 4. Comparison of MCC and Cohen's Kappa metrics for Multi-class classification models

CNN-BiLSTM-Attention (c)	CNN-LSTM model (b)	BiLSTM model (a)	Metric
0.969	0.924	0.884	MCC
0.970	0.924	0.883	Cohen's Kappa

These results provide strong evidence of the consistent relationship between the model's predictions and the actual classifications of data. This suggests that the proposed model has a high degree of stability and reliability when making decisions, particularly in environments where a high degree of reliability is required from a model with sufficient explanatory power to support the model's findings.

5.4. Multi-class Receiver Operating Characteristic (ROC) analysis

The Area Under the Curve (AUC) is one of the most common metrics used to measure the effectiveness of classification by the models. AUC values close to 1.0 indicate a strong ability of the model to distinguish between different classes. Based on the ROC curve analysis, the CNN-BiLSTM-Attention model outperformed all other compared models in detecting cyberattacks and distinguishing them from normal traffic. As illustrated in Fig. 3, the model showed exemplary performance through high AUC scores in most of the classification categories (e.g., backdoor, DDoS, DoS, MITM, ransomware, scanning, XSS, and normal). The AUC scores were also very high for the injection and password data categories (0.99), indicating that the CNN-BiLSTM-Attention model has a good ability to identify threats even in instances of low data overlap. The ROC curves are very close to the upper-left corner of the plot, indicating good recall rates and low false alarms when compared to random classification (AUC = 0.5).

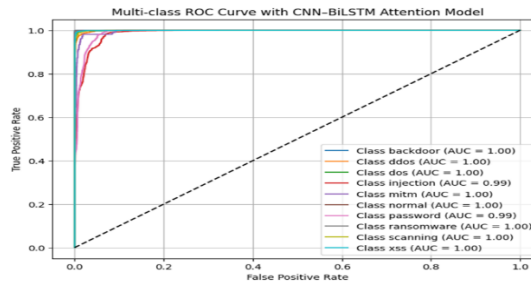


Fig. 3. Multi-class ROC curves for the CNN-BiLSTM-Attention classification model

This advantage is due to the combination of CNN layers for spatial feature extraction, BiLSTM for processing time sequences in both directions, and the attention mechanism for focusing on the most important characteristics. Performance was validated on independent test groups to ensure generalizability and reduce the likelihood of over-adaptation.

5.5. Binary classification performance

The binary classification results, illustrated in Fig. 4, show significant differences between the three models in discriminating between sensor faults and cyberattacks, which is the primary applied goal of this research in securing important IoT systems. The results will be discussed as follows:

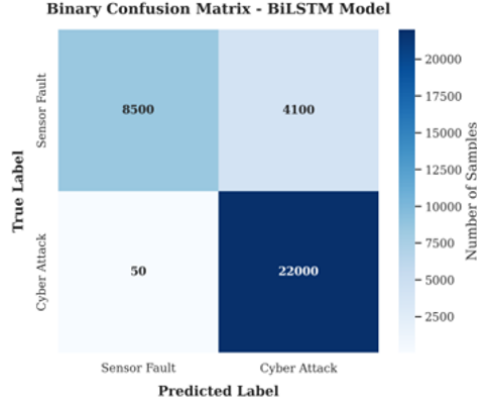
- The BiLSTM model (Fig. 4a) demonstrated high attack detection capability, with 22,000 true positives and 50 false negative cases, but it also recorded 4100 false positives due to misclassifying faults as attacks, resulting in a high False Alarm Rate (FAR) and indicating limited interpretive accuracy in distinguishing between faults and attacks.
- The CNN-LSTM model (Fig. 4b) was the best model for increasing the detection of attacks (32,011 true positives) and produced the highest number of false positive errors (4181) when classifying faults. This indicates that the CNN-LSTM model tends to generalize anomalous patterns without effective discrimination between causes of a fault.
- The proposed CNN-BiLSTM-Attention model (Fig. 4c) achieved very well-balanced results, achieving 31,726 True positives and 1945 false positives. This means that it has been able to accurately differentiate between cyberattacks and sensor failures and has improved model agility (explainability). This advantage is believed to be due to the ability of the attention mechanism to concentrate on precisely the right time-series and spatial patterns that differentiate a cyberattack from a sensor failure.

The research demonstrates that using a CNN with a BiLSTM and attention mechanism together creates a robust and interpretable system for identifying and classifying into two types of events: failure and attack. This is also indicative, given that there was a goal of producing an understandable deep learning system to classify either of these events.

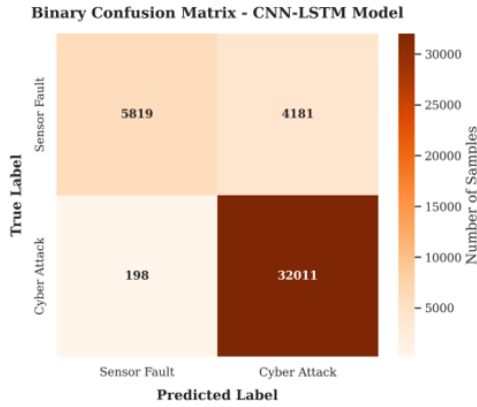
Using binary classification, only two classes of data were created for analysis: cyberattack and sensor failure. This correlates directly to generating an understandable deep learning method to identify and classify both types of events. A comparison of the performance of BiLSTM, CNN-LSTM, and CNN-BiLSTM-Attention models is displayed in Table 5.

With an accuracy of 94.25%, prediction precision of 94.22%, and an F1-score of 96.31%, the CNN-BiLSTM-Attention model exceeded all models as the best-performing model with the lowest false alarm rate (FAR = 19.45%). All models achieved an overall Detection Rate (DR) greater than 98%, but the CNN-BiLSTM-Attention model provided the best balance between accuracy and attack detection and false alarm rates. This level of superiority demonstrates the ability of the

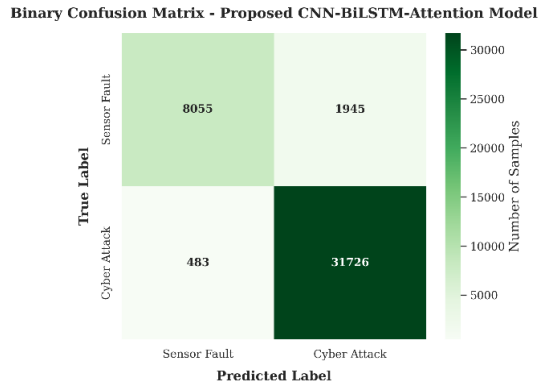
attention mechanism to identify distinct patterns of cyberattacks, thereby validating the applicability of the model for use in practice in identifying and securing critical IoT environments.



(a) Binary confusion matrix – The BiLSTM model



(b) Binary confusion matrix – The CNN-LSTM model



(c). Binary confusion matrix – The proposed CNN-BiLSTM-Attention model

Fig. 4. Comparison of binary confusion matrices for cyberattack and sensor fault classification

Table 5. Classification performance metrics of the proposed models in binary classification

Metric	BiLSTM	CNN-LSTM	CNN-BiLSTM-Attention
Accuracy	88.02%	89.62%	94.25%
Precision	84.29%	88.45%	94.22%
Recall Direction Rate (DR)	99.77%	99.39%	98.50%
F1-score	91.38%	93.61%	96.31%
FAR(%)	32.54%	41.81%	19.45%

To enhance the reliability of the evaluation for binary classifications, a set of statistically advanced measures was evaluated during this phase of the study (shown in Table 6). These included Matthews' Correlation Coefficient (MCC) & Cohen's Kappa. The CNN-BiLSTM-Attention model provided a higher MCC and Kappa value than the CNN-LSTM & BiLSTM models (MCC = 0.701, Kappa = 0.667; MCC = 0.751, Kappa = 0.722, respectively). These findings indicate strong levels of statistical agreement between predictions produced from each model and actual classifications, illustrating the stability and reliability of the proposed model for making decisions within high-risk settings requiring highly reliable models with strong explanatory power.

Table 6. Statistical performance metrics of the proposed models in binary classification

Metric	BiLSTM model	CNN-LSTM model	CNN-BiLSTM-Attention
MCC	0.751	0.701	0.837
Cohen's Kappa	0.722	0.667	0.832

5.6. ROC curve analysis

The proposed CNN-BiLSTM-Attention model demonstrated optimal performance when compared to the other models, as shown in Fig. 5, as evidenced by the highest AUC-score of 0.97, which makes it the most accurate and reliable model at distinguishing between cyberattacks and sensor failures. The second-best performing model was the CNN-LSTM model with an AUC-score of 0.91, while the BiLSTM model had the lowest AUC-score (0.85) of all three models.

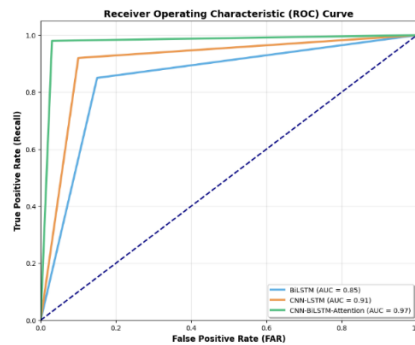


Fig. 5. ROC curves of the three models in the binary classification experiments

The ROC curve displays how effective the model was in predicting positive/untrue results by plotting the true positives (i.e., correctly identified positives) against the false positives (i.e., identified as positive when they're in fact negative). The green line representing the suggested model is much closer to the top left corner of the curve, which indicates that the model achieves the highest recall rate (the ability of a classification model to identify all relevant instances) and the lowest false positive rate (the ability of a model to predict a positive instance when it was in fact negative).

All models on the ROC are located above the line 0.5 (the diagonal line in the curve); therefore, all models have performed better than random classification (AUC = 0.5). Consequently, models created using this architectural classification method performed well.

The components of the new proposed model exhibit strength in terms of their effectiveness related to the provided models; the capability for CNNs to identify very complex spatially related patterns in images has been paired with the ability of BiLSTMs to identify the sequential nature of the data when viewed from two directions. The inclusion of an attention mechanism within the new model allows it to pay closer attention to only those taxonomically associated features and ignore any irrelevant information; therefore, the new model using CNN and BiLSTM can differentiate between relevant versus non-relevant taxonomic assignments at an AUC of 0.97 (this allows for use in the near “real time”).

5.7. The Performance and interpretability of the CNN–BiLSTM-attention model

An analytical framework was developed to enhance the transparency and trustworthiness of critical IoT systems by building on the inherent non-interpretability of deep learning systems using XAI. The framework is built on two complementary levels of explanation: global and local, using the SHAP value for global interpretability and an attention mechanism integrated into the model for local and temporal interpretability.

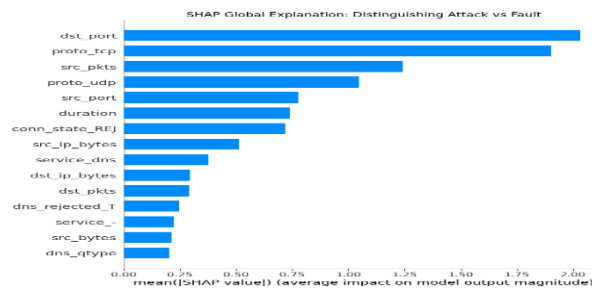


Fig. 6. Global SHAP feature importance for the CNN–BiLSTM-Attention model based on mean absolute SHAP values

Global interpretability using SHAP enables a clearer understanding of how the proposed model has been constructed to differentiate between cyberattacks over the entire dataset, as well as to enable the distinction to be made when either sensor failure or cyberattack has occurred. Fig. 6 presents the average absolute value of feature influences on the model output to identify the contribution of individual features to the overall model decision-making mechanism.

The results of this study revealed a clear dominance of network and transport layer features, with `dst_port` (destination port) being the most influential, followed by `proto_tcp` and `src_pkts`. This learning behavior is consistent with the theoretical foundations of cybersecurity in critical environments, as these characteristics are key structural indicators of systematic attack patterns, such as network scanning and Distributed Denial-of-Service (DDoS) attacks, which typically target specific ports and protocols within the network infrastructure.

These results show that the proposed model not only achieves great classification performance but also has a high level of interpretability, as its decisions are based on clearly relevant security properties rather than random statistical correlations. The inclusion of SHAP improves the model's openness and confidence in its outcomes, which is vital for the use of AI technologies in critical IoT systems that demand high reliability and clear interpretive support for real-time decision-making.

To achieve interpretability at the individual sample level and to understand how the model makes decisions at various times, the proposed model's attention weight distributions were examined. This study allows you to follow the contribution of different attributes within the data's time sequence and get a complete perspective of the model's behavior during the classification process.



Fig. 7. Attention weight distribution for features at the sample level (Sample ID: 10). The red bars indicate the most influential features attended by the model for making the classification decision, reflecting local interpretability

Fig. 7 demonstrates a graphical representation of the attention weights for Sample ID 10. As presented, there are various peaks, as shown in red, indicating the

time steps that were given the closest attention during the classification process. As shown, some of these peaks are as high as 0.75. These peaks are related to critical trigger points for abnormal system behavior, as well as potential malfunctions in sensors. The attention mechanism helps to identify situations where there are major behavioral changes, rather than relying on averages.

5.8. Feature impact analysis (bee-swarm plot)

To conduct a more comprehensive evaluation of model interpretability, a bee-swarm plot was created within the SHAP summary to find out how different features contribute to the model's output via their respective directions (i.e., increasing and decreasing). Understanding the directional impact that a feature has on model decisions concerning detecting (or not detecting) cyberattacks and sensor faults can be accomplished using this type of analysis.

SHAP values in Fig. 8 show each input feature's impact on model predictions. DNS attributes like `dns_RA` and `dns_rejected`, marked in red, greatly influence classification. The findings suggest DNS data is useful for spotting cyberattacks, clarifying model decisions. SHAP analysis explains how specific feature values affect individual predictions, both generally and specifically. Unusual DNS patterns, like spoofing and DDoS attacks, often indicate malicious activity, confirming these features' relevance in detecting cyberattacks.

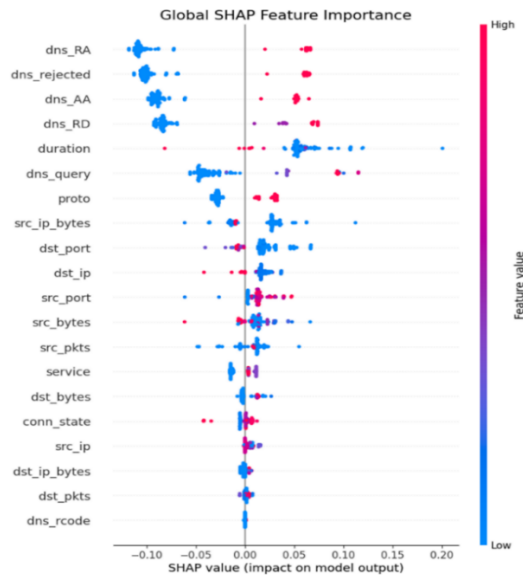


Fig. 8. SHAP feature contributions for cyber attack and sensor fault classification

5.9. Discussion

The evaluation results of the CNN-BiLSTM-Attention model using the TON_IoT dataset have shown that the proposed model outperformed the BiLSTM and CNN-LSTM models in both multi-class and binary classifications, achieving the highest scores across all evaluation metrics, including accuracy, F1-score, MCC, Cohen's Kappa, and AUC. The proposed model's consistent performance across numerous metrics demonstrates that its effectiveness is not reflected by a single metric but rather by a comprehensive improvement across multiple dimensions of classification quality.

The proposed hybrid architecture has produced superior performance because BiLSTM learns bidirectional dependencies while CNN helps extract local feature patterns. Nevertheless, when used independently, these models may encounter problems with handling overlapping attacks. Although the CNN-LSTM model improves overall representation, it may still suffer from feature confusion. Introducing the attention mechanism enables the model to focus on the most relevant features while suppressing less informative ones, thereby reducing the false alarm rate.

The high values of MCC and Cohen's Kappa confirm the robustness of the proposed model under complex and imbalanced conditions. The high value of AUC proves the efficiency and ability of the proposed algorithm to recognize both normal and attack conditions across a wide range of decision thresholds.

The interpretability analysis using SHAP and attention mechanisms proves the reliance of the proposed model on appropriate security features (such as `dst_port` and `proto_tcp`), which improves the efficiency of this model in IoT systems. Finally, the research results demonstrate the achievement of a satisfactory balance between the accuracy, robustness, and interpretability of the proposed model.

6. Conclusions

This study proposed a hybrid deep learning model, called CNN-BiLSTM-Attention, to enhance the security of critical IoT systems by accurately distinguishing between cyberattacks and sensor failures. The model was trained and evaluated using a realistic TON-IoT dataset designed to simulate IoT networks. The experimental results showed the superiority of the proposed model over traditional models in terms of accuracy and balance between retrieval, classification accuracy, and reduction of false alarm rate in both multi-class and binary classification. ROC curve analyses show a high AUC value, confirming that the proposed model has a high ability to distinguish between cyberattacks, normal behavior, and sensor faults.

The model also demonstrated high stability and reliability, supported by advanced statistical metrics such as MCC and Cohen's Kappa. Furthermore, the integration of SHAP and attention mechanism technologies with the proposed model has provided global and local interpretation, enhanced transparency, and facilitated a clear understanding of the model's decision-making process. Overall, the results show that the proposed model achieves high performance and an advanced level of interpretability, making it appropriate for application in critical IoT systems.

References

1. Moss, J., G. Wordon, Y. Duclos, Q. Liu, Q. Wang, J. Wang. Explainable AI in IoT: A Survey of Challenges, Advancements, and Pathways to Trustworthy Automation. – *Electronics*, Vol. **14**, 2025, No 23, 4622.
2. Suryateja, P. S., K. V. Rao. A Survey on Lightweight Cryptographic Algorithms in IoT. – *Cybernetics and Information Technologies*, Vol. **24**, 2024, No 1, pp. 21-34.
3. Sebestyen, H., D. E. Popescu, R. D. Zamanda. A Literature Review on Security in the Internet of Things: Identifying and Analysing Critical Categories. – *Computers*, Vol. **14**, 2025, No 2, 61.
4. Elsayed, Z., A. Abdelgawad, N. Elsayed. Cybersecurity and Frequent Cyber Attacks on IoT Devices in Healthcare: Issues and Solutions. – *arXiv Preprint*, 2025.
5. Tsimenidis, S., T. Lagkas, K. Rantos. Deep Learning in IoT Intrusion Detection. – *Journal of Network and Systems Management*, Vol. **30**, 2022, No 1, 8.
6. Sharma, S., K. Gupta, D. Gupta, S. Rani, G. Dhiman. An Insight Survey on Sensor Errors and Fault Detection Techniques in Smart Spaces. – *Computer Modeling in Engineering & Sciences*, Vol. **138**, 2024, No 3, 2029.
7. Khraisat, A., A. Alazab. A Critical Review of Intrusion Detection Systems in the Internet of Things: Techniques, Deployment Strategy, Validation Strategy, Attacks, Public Datasets and Challenges. – *Cybersecurity*, Vol. **4**, 2021, 18.
8. Sarker, I. H., A. Kayes, S. Badsha, H. Alqahtani, P. Watters, A. Ng. Cybersecurity Data Science: An Overview from Machine Learning Perspective. – *Journal of Big Data*, Vol. **7**, 2020, No 1, pp. 1-29.
9. Imran, Y., Y. Xiang, L. Ali, A. S. Noor, K. Sarpong. CNN-GRU-FF: A Double-Layer Feature Fusion-Based Network Intrusion Detection System Using Convolutional Neural Network and Gated Recurrent Units. – *Complex & Intelligent Systems*, Vol. **10**, 2024, pp. 1-18.
10. Mallampati, S. B., H. Setha. Enhancing Intrusion Detection with Explainable AI: A Transparent Approach to Network Security. – *Cybernetics and Information Technologies*, Vol. **24**, 2024, No 1, pp. 98-117.
11. Hassija, V., V. Chamola, A. Mahapatra, A. Singal, D. Goel, K. Huang, A. Hussain. Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. – *Cognitive Computation*, Vol. **16**, 2024, No 1, pp. 45-74.
12. Ribeiro, M. T., S. Singh, C. Guestrin. Why Should I Trust You? Explaining the Predictions of Any Classifier. – In: *Proc. ACM SIGKDD*, 2016.
13. Lundberg, S. M., S. I. Lee. A Unified Approach to Interpreting Model Predictions. – In: *Advances in Neural Information Processing Systems*, 2017, pp. 4765-4774.

14. V i m b i, V., N. S h a f f i, M. M a h m u d. Interpreting Artificial Intelligence Models: A Systematic Review on the Application of LIME and SHAP in Alzheimer's Disease Detection. – Brain Informatics, Vol. **11**, 2024, No 1, 10.
15. S a d h w a n i, S., M. A. H. K h a n, R. M u t h a l a g u, P. M. P a w a r. A Hybrid BiLSTM-CNN Approach for Intrusion Detection for IoT Applications. – Scientific Reports, Vol. **16**, 2025, No 1, 155.
16. Z h a n g, C., J. L i, N. W a n g, D. Z h a n g. Research on Intrusion Detection Method Based on Transformer and CNN-BiLSTM in Internet of Things. – Sensors, Vol. **25**, 2025, No 9, 2725.
17. Y a r a s, S., M. D e n e r. IoT-Based Intrusion Detection System Using New Hybrid Deep Learning Algorithm. – Electronics, Vol. **13**, 2024, No 6, 1053.
18. C u i, B., Y. C h a i, Z. Y a n g, K. L i. Intrusion Detection in IoT Using Deep Residual Networks with Attention Mechanisms. – Future Internet, Vol. **16**, 2024, No 7, 255.
19. S a i y e d, M. F., I. A l - a n b a g i, M. S. H o s s a i n. An Explainable Deep Learning System for Cyberattack Detection in Internet of Energy Networks. – IEEE Network, Vol. **40**, 2026, No 2, pp. 39-46.
20. S h t a y a t, M. M., M. K. H a s a n, R. S u l a i m a n, S. I s l a m, A. U. R. K h a n. An Explainable Ensemble Deep Learning Approach for Intrusion Detection in Industrial Internet of Things. – IEEE Access, Vol. **11**, 2023, pp. 115047-115061.
21. B h o w m i k, A. TON_IoT Network Dataset. Kaggle, 2025.
22. S h a r m a, K. P., T. N a g p a l, T. V o r a, A. Y a d a v, M. I. A b d u l l a h, B. J a y a p r a k a s h, B. B. B u k a t e. Interpretable Intrusion Detection for IoT Environments Using a Self-Attention-Based Explainable AI Framework. – Scientific Reports, Vol. **15**, 2025, No 1, 39937.
23. K o p e t z, H., W. S t e i n e r. Internet of Things. Springer, 2022, pp. 325-341.
24. E l h a n a s h i, A., P. D i n i, S. S a p o n a r a, Q. Z h e n g. Integration of Deep Learning into the IoT: A Survey of Techniques and Challenges for Real-World Applications. – Electronics, Vol. **12**, 2023, No 24, 4925.
25. T h a k u r, D., J. K. S a i n i, S. S r i n i v a s a n. DeepThink IoT: The Strength of Deep Learning in Internet of Things. – Artificial Intelligence Review, Vol. **56**, 2023, No 12, pp. 14663-14730.
26. C h i n n a s a m y, R., M. S u b r a m a n i a n, S. V. E a s w a r a m o o r t h y. Deep Learning-Driven Methods for Network-Based Intrusion Detection Systems: A Systematic Review. – ICT Express, Vol. **11**, 2025, No 1, pp. 181-215.
27. L i, B., J. L i, M. J i a. ADFCNN-BiLSTM: A Deep Neural Network Based on Attention and Deformable Convolution for Network Intrusion Detection. – Sensors, Vol. **25**, 2025, No 5, 1382.
28. K o w s h e r, M., A. T a h a b i l d e r, M. Z. I. S a n j i d, N. J. P r o t t a s h a, M. S. U d d i n, M. A. H o s s a i n, M. A. K. J i l a n i. LSTM-ANN & BiLSTM-ANN: Hybrid Deep Learning Models for Enhanced Classification Accuracy. – Procedia Computer Science, Vol. **193**, 2021, pp. 131-140.
29. N a t h a, S., F. A h m e d, M. S i r a j, M. L a g a r i, M. A l t a m i m i, A. A. C h a n d i o. Deep BiLSTM Attention Model for Spatial and Temporal Anomaly Detection in Video Surveillance. – Sensors, Vol. **25**, 2025, No 1, 251.
30. H a s s o u n e h, N., S. A l - s h a r a e h. Intrusion Detection in IoT Networks Using LSTM Deep Learning Models with the UNSW-NB15 Dataset. – Proc. ICTCS, IEEE, 2025, pp. 263-269.

31. Gupta, H., A. Jadhav, A. S. Bishit. Comparative Analysis of Machine and Deep Learning Models for Intrusion Detection in Fog-Enabled IoT Networks. – International Journal of Networked and Distributed Computing, Vol. **14**, 2026, No 1.
32. Abualhaj, M., H. Adeeel, K. Masood, H. Soltani, H. Zemmouri, M. M. R. Aly, S. Mehmood. Comparative Analysis of LSTM-Based Variant Models for Detecting Attacks in IoT Networks. – Journal of Computing & Biomedical Informatics, Vol. **10**, 2025, No 1.
33. Tarek, R., A. Elshenawy, M. I. Assady, M. A. Madkour. Automated Diagnosis of Dental Diseases Using Deep Learning on Radiographic Images. – SN Computer Science, Vol. **6**, 2025, No 6, 751.
34. Sinha, J., M. Manollas. Efficient Deep CNN-BiLSTM Model for Network Intrusion Detection. – Proc. ICAIPR, ACM, 2020, pp. 223-231.
35. Abduljabbar, R. L., H. Dia, P. W. Tsai. Development and Evaluation of Bidirectional LSTM Freeway Traffic Forecasting Models Using Simulation Data. – Scientific Reports, Vol. **11**, 2021, No 1, 23899.
36. Brokwa-larbi, K. The Role of IoT and XAI Convergence in the Prediction, Explanation, and Decision of Customer Perceived Value (CPV) in SMEs: A Theoretical Framework and Research Proposition Perspective. – Discover Internet of Things, Vol. **5**, 2025, No 1, 4.
37. Dwivedi, R., D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel. Explainable AI (XAI): Core Ideas, Techniques, and Solutions. – ACM Computing Surveys, Vol. **55**, 2023, No 9.
38. Wang, Y. A Comparative Analysis of Model-Agnostic Techniques for Explainable Artificial Intelligence. – Research Reports on Computer Science, 2024, pp. 25-33.

*Received: 01.04.2026; First Revision: 04.06.2026; Second revision: 10.06.2026;
Accepted: 12.06.2026*

© 2026 by the authors. Licensee IICT-BAS, Sofia, Bulgaria. This article is an open access article distributed under the terms and conditions of the **Creative Commons Attribution-NonCommercial-NoDerivatives** license.