Trust-Aware Deep Learning Framework for Mitigating False Data Injection and Integrity Breaches in IoT Ecosystems

OPEN ACCESS

Volume: 13

Special Issue: 3

Month: February

Year: 2026

P-ISSN: 2321-788X

E-ISSN: 2582-0397

Citation:
Devi, V., S. Vijaya, et al. "Trust-Aware Deep Learning Framework for Mitigating False Data Injection and Integrity Breaches in IoT Ecosystems." *Shanlax International Journal of Arts, Science and Humanities*, vol. 13, no. 3, 2026, pp. 97–106.

DOI:
https://doiorg/10.34293/sijash.v13iS3-i1-Feb.10267

Dr. V. Devi
Professor, PG & Research Department of Computer Science
Thiruthangal Nadar College, India

Mrs. S. Vijaya
Research Scholar, PG & Research Department of Computer Science
Thiruthangal Nadar College, India

Dr. A. Ambeth Raja
Associate Professor, PG & Research Department of Computer Science
Thiruthangal Nadar College, India

S. Sathya
Department of Computer Applications
Thiruthangal Nadar College, Chennai

Abstract

IoT systems grow explosively so they are becoming more vulnerable to False data injection (FDI) and data integrity attacks. This can impact the availability and reliability of the system. Standard security mechanisms are insufficient to provide the appropriate security level required for large, dynamic, heterogeneous environments. In this paper we propose a Trust Aware Deep Learning Framework (TADLF) for intelligent detection and prevention of integrity, related cyber-attacks for IoT. We propose a Dynamic Trust Assessment (DTA) component in order to calculate trust coefficient scores by determining device trustworthiness using historical data transmission behavior, source authenticity, and device communication records. In addition, an attention, based hybrid CNN BiLSTM module is proposed to effectively learn the spatiotemporal anomalies. By combining the DTA and an extended feature set extracted based on the trust score, the attention, based hybrid model is achieved average accuracy of 98.4%, precision of 98.1%, recall of 98.6% and F1 score of 98.3% with 1.5% false positive alarm rate, IoT dataset having 461,000 data records, for FDI attacks compared to base models (RF, XGBoost and LSTM).

Keywords: Internet of Things (IoT), Trust, CNN, LSTM

Introduction

This evolving revolution, the Internet of Things (IoT), is the current digit environment revolution and provides great amount of network devices and systems, actuators, sensors, intelligent devices etc for variety of applications with significant impact on many sectors of economy (smart city, health monitoring, industrial automation,

energy systems, intelligent transportation etc). Through enormous complex interconnected systems providing real-time information and automatic control, their efficiency and quality of service can be significantly improved [1, 4]. However, the huge deployment of resource scarce systems with disparate protocols and distributed architecture brings forward some security concern.

FDI attack and decreasing data integrity are the top threats in the IoT world, which attackers injected malicious data in the measured or transmitted sensor data in order to affect the operation of system, create false response of the system and conceal their evil work [5, 7]. Different from the common DoS or brute force attacks, integrity attacks could be stealth, run on background quietly for a long time and cause huge functional failures. This could be the serious disaster such as power stability loss in smart grid, misdiagnosis in healthcare IoT, shutdown in industrial control system and so on [8, 10].

Traditional crypto and rule-based IDSs are deployed since quite long, in order to ensure the security of the IoT communications. Although its use has been able to assure integrity of information, the security of crypto/authentication means could be lost through insider attack, infected nodes or several evasion techniques that correctly manipulate legitimate access credentials. Signature, based detection systems fail to provide with kinetics and zero-day, provenance of the integrity attacks. All these fact highlights the need of building intelligent security systems that learn to adapt, inside behavior of the system in tracking down anomalous behavior in large IoT data[11, 12].

Recently, machine learning, deep learning methods have demonstrated a great potential in intrusion detection, pattern detection and cyber threat intelligence. Deep learning methods based on CNNs, RNN, LSTMs, recently transformers are successfully adapted in the areas of intrusion detection and network traffic analysis [13,15]. However, deep learning models mostly focus on the raw features or statistics parameters of traffic, ignoring the ever, changing trustworthiness of IoT devices. So, such deep learning models may not provide reasonable robustness in correlated attacks, malicious compromised nodes, aging malicious data.

Authentication management is one effective mechanism to calculate node‘s behavior history over time including its past interactions, data consistency and trust characteristics in a distributed IoT environment. It is also able to calculate the device‘s trustworthiness as a figure to distinguish normal and malicious nodes once the existing security mechanisms are defeated. Currently, trust based security mechanisms for IoT have only used heuristic scoring and basic machine learning techniques, therefore limited in scalability and accuracy of attack detection.

This paper proposes a novel modeling framework named as Trust Aware Deep Learning Framework(TADLF) to tackle these issues that integrates dynamically time, varying node trustworthiness calculation into a hybrid deep learning modeling framework, where the behavior based trustworthiness scores are simultaneously modeled with spatiotemporal anomaly indices. The trust scores are dynamically updated based on communication history of a node, deviations of side information, and past trustworthiness and then incorporated into the deep learning model as weighted intelligence features. The proposed model shows higher absorbency to hidden data manipulations and favorable run, to, run adaptability, and an on, line trigger system is designed for the prompt isolation of malignant nodes for the further enhanced system robustness.

The major contributions of this work are summarized as follows:

- A new trust-aware deep learning framework which jointly models device trust and anomaly behavior to detect integrity threat in IoT networks
- A dynamic trust evaluation mechanism that considers both the long-term stability and the recent anomalies in IoT data streams.

- Hybrid CNN-BiLSTM model is developed for the efficient spatial-temporal feature learning of false data injection patterns.
- The extensive experimental validation proved accuracy and robustness and lower false alarm rate compared to the conventional ML/DL methods..

The remainder of this paper is organized as follows: Section 2 reviews related work on FDI detection and IoT data integrity protection. Section 3 presents the proposed trust-aware deep learning framework and system architecture. Section 4 describes the experimental setup and performance evaluation and Section 5 concludes the paper with future research directions.

Literature Review

FDI (False Data Injection) and Data Integrity attacks have become essential security issues in the cyber-physical systems and in the IoT based systems especially in the areas of smart grids and IoT based healthcare. The state-of-the-art studies have been adopting deep learning and reinforcement learning framework for addressing the sly and adaptable behavior of such attacks.

Huang et al. Developed an attention-aware deep reinforcement learning framework for FDI attacks detection in smart grid systems [1]. It combined attention mechanism with deep Q-networks, which dynamically emphasized relevant states and could perform better detection under sophisticated attack circumstances. In terms of effectiveness, the proposed algorithm also demonstrated better adaptivity comparing to common anomaly detection methods. Kurt et al. Also proposed a reinforcement learning based online cyber-attack detection scheme in smart grid where intelligent agents could effectively learn optimal defense strategies to perform timely detection on malicious data manipulation [2]. Reinforcement learning has good adaptivity but it has high training cost and is only applied to grid control scenarios. Therefore, its scalability in the heterogeneous IoT environment is not satisfactory.

To scale data driven resiliency techniques, Chen et al designed a data-driven automatic generation control mechanism that incorporates machine learning to deal with FDI attacks on industrial energy systems robustness [3]. Their system focuses on stable operation of systems given corrupted measurements. On a parallel front, Srinivasan et al designed a multi-label deep learning classifier to handle multiple types of FDI attacks concurrently. They also showed that the attack differentiation ability on the communication networks of a smart grid can be improved [4]. However, those systems only worked well on well-structured energy systems and used feature engineering specific to the systems of interest without incorporating any behavior trust model.

Besides energy infrastructures, other kinds of integrity protection of IoT-enabled healthcare systems have been extensively studied. Hussein et al. Provided a deep learning applications perspective to smart health applications in an IoT environment where anomaly detection and security monitoring are key for patient data security [5]. Naik et al. Developed a hybrid deep learning framework for enhanced security and integrity of healthcare IoT, using a combination of CNNs and RNNs for identifying complicated attacks [6]. Algethami and Alshamrani proposed a deep learning-based cybersecurity framework for the IoHT, which aimed at real-time threat detection and response of intrusions [7]. All these researches proved that the models could identify integrity threats to healthcare systems, yet they all treated the IoT devices as trustworthy without considering dynamically.

From a signal processing perspective, Yadav and Pradhan developed a wavelet probability distribution mapping technique for detecting and fixing dynamic data injection attacks in wide-area monitoring systems [8]. This approach was capable of identifying the transient anomaly induced by the malicious alteration and resulted in better detection of dynamic attack behaviors. However, wavelet-based methods demand skillful selection of parameters and are not efficient for large and heterogeneous IoT networks with massive high-dimensional data streams.

Yang and Zhang provide a more global overview by performing a thorough literature survey of anomaly detection schemes in IoT environments and demonstrating the migration from the initial statistical and rule-based solutions toward the recent use of deep learning and hybrid intelligent models [9]. They mention the long-term problems that persist such as concept drift, heterogeneous devices, data imbalance and no trust models and argue that the system needs to be able to intelligently learn dynamic attacks, along with a reliability measure of the individual distributed devices.

Recently, research also starts to consider the adversarial intelligence in integrity attack. The paper by Xiao et al. Studied the reinforcement learning based false data injection attack strategies in smart grids and illustrated how intelligent attackers can adapt to bypass traditional defense [10]. It further proves that adaptive and behavior-oriented defense schemes against dynamically changing attacker strategies are essential..

Research Gaps

Despite significant progress, existing studies reveal:

- Heavy domain dependence (smart grids vs general IoT)
- Lack of trust-aware behavioral modeling
- Weak robustness against adaptive FDI attacks
- Scalability constraints in RL and signal-based approaches

Methodology

This section proposes a Trust-Aware Deep Learning Framework (TADLF) for detecting and preventing false data injection and data integrity attacks in IoT systems. The proposed framework combines dynamic trust modeling and a hybrid deep learning architecture to improve resistance to furtive and evolving integrity threats.

To demonstrate the validity of the proposed framework the ToN-IoT (Telemetry of Network IoT) dataset is utilized. This dataset is a freely available dataset and it is accessible from University of New South Wales (UNSW Canberra Cyber Range Lab). In Table 1 it can be seen the parameters.

Table 1 Dataset Parameters

Parameter	Description
Total records	~461,000
Attack types	FDI, data manipulation, spoofing, injection, malware
Features	44 network & telemetry attributes
Classes	Normal + multiple attack categories
Data format	CSV

Overall System Architecture

The proposed TADLF follows a multi-layer security architecture consisting of:

1. IoT Data Acquisition Layer
2. Preprocessing and Feature Engineering Layer
3. Trust Evaluation Layer
4. Deep Learning Detection Layer
5. Decision and Mitigation Layer

This work introduces the Trust-Aware Deep Learning Framework (TADLF), a security architecture based on multiple layers enabling secure and efficient detection and prevention of False Data Injection and data integrity attacks in IoT environments. This multi-layered design

promotes flexibility, scalability and real-time response by seamlessly integrating behavioral trust information into deep learning based anomaly detection mechanisms.

IoT Data Acquisition Layer: This layer acts as the foremost interface between the physical devices and the security system. This layer accumulates the heterogeneous data streams originating from the sensors, actuators and smart nodes deployed across the IoT network. Examples of data streams may be telemetry metrics, network attributes, devices states, sensor readings etc. sent using MQTT, CoAP, HTTP protocols. Owing to resource limited and distributed nature of IoT, this layer must be located at the edge of the network to capture low-latency data and perform initial filtering. Real-time continuous stream of data allows detecting anomaly at early stage before it could cause system-wide disturbance.

After acquiring data, the Preprocessing and Feature Engineering Layer converts the raw IoT data into a format ready for ML algorithms. The stage consists of de-noising the data, dealing with missing data points, normalizing the data, and encoding the categorical attributes. Temporal windowing approaches were employed to maintain sequential dependencies, necessary in detecting slow and/or covert FDI attacks. Furthermore, statistical features such as data deviation, entropy variation, packet transmission rate, and sensor fluctuation were calculated to improve anomaly detection sensitivity. This processing stage makes the data more uniform, reduces the dimensionality and further increases the efficiency and convergence speed of deep learning algorithms.

The Trust Evaluation Layer brings behavioral intelligence into the security infrastructure through its real-time evaluation of the trustworthiness of every IoT node. The trust score is calculated from node data consistency, communication reliability, and past behavior integrity patterns. Nodes that report inconsistencies often, transmit abnormally often, or report inconsistently gradually become less trusted. Through constant updates in real-time, this layer is able to distinguish from legitimate yet noisy devices and malicious or compromise devices. The trust vector is then incorporated into the feature set, which in turn is used by the detection model.

The fundamental of TADLF is the Deep Learning Detection Layer, a hybrid CNN-BiLSTM structure capable of capturing both spatial and temporal attack features. CNN extracts high-level spatial correlations in multidimensional IoT data and thus discovers minute manipulations in such data that are frequently the feature of false data injection attacks. BiLSTM, on the other hand, effectively captures the temporal correlation, in other words learning evolution patterns of attacks and long-term behavioral deviations. This hybrid network structure thus leads to better detection performance in terms of detecting both real-time anomaly and the historical anomaly. It can thus be immune to subtle and slow moving integrity attack.

The last component of the Decision & Mitigation layer processes the classification decisions and implements appropriate response strategies. On identifying integrity threat, the system can initiate node isolation mechanisms, block malicious traffic, adjust trust score, and notify network administrators, which automates response and stops the attack from spreading. Adaptive thresholding is another valuable feature to detect emergent threats while restricting the number of false alarms.

Data Preprocessing

Preprocessing plays a key role because it can directly determine how efficient and successful the machine learning based IDS in IoT. The initial telemetry dataset of the Internet of Things, ubiquitous, heterogeneous, volume and so forth, needs to be processed and formatted into a normalized one to train the model.

In first instance data cleaning procedures have been performed. Data cleaning intends to remove the dirty, duplicated and incomplete data entries. The missing values have been impetrated using statistical imputation methods for filling the empty slots among the others. Duplicate data entries,

often created by duplicate sending, and/or errors in logging, have been removed in order to avoid the bias to the data, and possible overfitting of the machine learning models.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

After that, in order to normalize numerical features, the Min-Max normalization is implemented. It rescales the features [0, 1]. The scaling is needed in order that every feature have the same contribution to the cost function in gradient descent and that none of the features overshadow any other. The transformation is defined as:

where X represents the original feature value, and and denote the minimum and maximum values of the corresponding feature, respectively.

The one-hot encoding is utilized to handle the categorical attributes, e. g., type of protocol and class of device. It transforms these categorical attributes into binary vectors. The deep learning model can therefore be able to recognize the categorical information.

The number of dimensions is reduced and unnecessary information is filtered out through feature selection and extraction methods. Using PCA (Principal Component Analysis) the high dimensional data is projected onto a low dimension space which would preserve most of the variance. Also, the features relevant for classification of an attack are selected using mutual information-based ranking to increase speed and detection rate.

Due to the presence of the imbalanced ratio between normal traffic samples and attack samples, we have utilized the mitigation techniques to deal with class imbalance issue in IoT datasets. We adopt SMOTE (Synthetic Minority Oversampling Technique) to create synthetic attack samples so as to balance the distribution of both normal and attack classes. Class weighting is employed while training the model to penalize misclassification of minority attack classes so that to lower down the rate of false negatives.

Last but not least, the temporal relations are kept through the sliding window approach of segmentation that group together the successive samples in fixed length segments, so that the deep learning model could identify progressive attack behavior or sustained integrity failure.

Trust Evaluation Model

In order to be robust against covert falsification and integrity-based attacks, TADLF integrates a dynamic trust model to measure node behavior credibility of individual IoT devices. In contrast to the standard anomaly detection systems, which solely consider the statistical deviation, the trust model provides intelligent context information by dynamically tracking node trustworthiness during the operational periods.

Each IoT device is given a trust score that is based on three main behavioral features: Data Consistency (DC), Communication Reliability (CR), and Historical Integrity (HI). DC is used to estimate how far the value sensed is away from expected operational values or neighbors' values and thus to detect the corrupted/generated data. CR is used to detect how stable the packet delivery, sending frequency and connection duration of each node is. HI measures the behavioral history, which is used to estimate the behavior in the long term based on history of attacking participation, the rate of anomaly and long-term deviation from the norm.

The overall trust score for node iii is computed as a weighted aggregation of these indicators:

$$T_i = \alpha DCi + \beta CR_i + \gamma HI_i$$

Subject to the constraint:

$$\alpha + \beta + \gamma = 1$$

where α, β, and γ represent weighting coefficients that control the relative influence of each behavioral component. In this study, empirically optimized values of α=0.4, β=0.35, and γ=0.25 are adopted to prioritize data credibility while preserving communication stability and historical behavior.

The mechanism to promote adaptability of trust, nodes demonstrating persistent deviation or anomalous behavior are gradually penalized in trust. This can detect compromised insiders or hidden attackers quickly, and also avoid a transient noisy node being over-penalized

The calculated trust values are then injected into the feature vector, in the form of weighted intelligence features, and passed into the deep learning detection module. Through the combination of behavior-based trust information with spatial-temporal anomaly features, greater sensitivity to small integrity attacks and higher resilience to adaptive attack techniques are gained.

In sum, the trust evaluation layer is an context based reliability filter that enhances the decision-making ability of the CNN-BiLSTM hybrid model to take proactive measure and mitigate the false data injection attacks in heterogeneous IoT environment.

Deep Learning Detection Module

At the center of the suggested Trust-Aware Deep Learning Framework (TADLF) is a hybrid CNN-BiLSTM which can learn efficiently spatial correlations and temporal dependencies for IoT data streams under false data injection and integrity attacks.

CNN Layer (Spatial Feature Extraction)

CNN module: the CNN module act as a spatial feature extractor. It is used to learn the high-level feature from the multidimensional IoT telemetry inputs. Two convolutional layers with 3x3 kernels size were used to extract local pattern in feature dimension. The first layer uses 64 filters, and the second layer uses 128 filters to gradually learn complex patterns of anomalies. The rectifying linear unit (ReLU) activation function was used to non-linearly transform, and it help to speed up the training, defined as:

$$f(x) = max(0,x)$$

This activation can prevent negative reactions while still having informative positive feature activations that represent abnormal data behavior.

BiLSTM Layer (Temporal Dependency Learning)

After spatial features extraction, the produced feature maps are fed into BiLSTM layer which learn spatio-temporal dependence in sequential IoT data. While an LSTM can only learn time dependencies in one temporal direction, BiLSTM learn information from both forward and backward directions, allowing learning the entire temporal context. The hidden state transition can be modeled as:

$$h_t = BiLSTM(x_t, h_{t-1})$$

By enabling two-way temporal modeling, the detection engine can discover evolving patterns of attack, slow progress of integrity violation, or the long-term deviation in behavior associated with stealthy FDI attacks.

The learned spatiotemporal representations are then fed into the fully connected layer of 128 nodes for merging the extracted features into a low-dimensional decision space. After the fully connected layer, the output layer uses the Softmax activation function to classify normal behaviors and several kinds of integrity attacks.

Hyperparameter Tuning Strategy

Hyperparameter optimization is critical to achieve high detection accuracy and model generalization. The following strategy is applied:

1. Learning Rate (η) – Grid search is performed within [0.0001, 0.01]. Optimal value: 0.001, ensuring stable convergence without overshooting.
2. Batch Size – Evaluated between 32, 64, and 128. Optimal batch size: 64, balancing memory efficiency and gradient stability.
3. Number of Epochs – Determined via early stopping on validation loss. Maximum set to 50 epochs to prevent overfitting.
4. Dropout Rate – Tested in [0.2, 0.5] to reduce overfitting; final dropout: 0.3 applied after BiLSTM.
5. Number of CNN Filters – Tuned through cross-validation: 64 and 128 filters selected for first and second layers, respectively.
6. BiLSTM Hidden Units – Tuned in [64, 128, 256]; 128 hidden units provided the best trade-off between accuracy and computational cost.
7. Optimizer Selection – Adam optimizer chosen for adaptive learning rate adjustments and faster convergence.

Experimental Setup and Results

For the experimental assessment of the proposed Trust-Aware Deep Learning Framework (TADLF), a robust hardware setup was utilized that supports parallel processing for the large volume of data from the IoT environment and enables the real-time model training. The system configuration included an Intel Core i7-12700 CPU, 32 GB of RAM and NVIDIA RTX 3080 GPU. The hardware offers adequate capacity to handle deep learning tasks and support fast, parallelized matrix multiplication, required for the CNN-BiLSTM model. Python 3.11 was used as the base software language with TensorFlow 2.12 and Keras used as the core deep learning libraries for easy model implementation, GPU acceleration and optimized training loops. Libraries including Scikit-learn 1.3, NumPy and Pandas were used for feature engineering, data preprocessing, performance evaluation and statistical calculations.

The ToN-IoT dataset of the UNSW Canberra Cyber Range Lab has been employed in this empirical study. It constitutes a well-rounded IoT telemetry dataset which comprises 461,000 records of 44 attributes describing the measurements, network traffic and behaviors of IoT devices. It comprises normal operational data along with falsified data injection (FDI), data manipulation, spoofing and injection attacks. In turn, it constitutes a realistic heterogeneous environment for measuring performance with respect to variety of integrity attacks. An 80:20 train-test split strategy was employed in order to ensure adequate generalization of the model and to validate the framework against data which had never been encountered before, such that results are representative of those for large-scale IoT environments.

Various researches suggested evaluating IoT security frameworks on the basis of multi-metric performance analysis, measuring detection rate, confidence and false alarm rate under varied attack schemes [1], [2]. According to that, the presented TADLF framework was experimented on ToN-IoT dataset and the performance with respect to detection for varied types of attacks is presented in Table 2 below. The metrics measured were Accuracy, Precision, Recall, F1-Score and False Alarm Rate (FAR).

Table 2: Detection Performance of TADLF Across Attack Types

Attack Type	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	FAR (%)
Normal	99.2	99.3	98.9	99.1	0.8
FDI	98.7	98.5	99.0	98.7	1.2
Data Manipulation	98.4	98.0	98.9	98.4	1.5
Spoofing	97.9	97.5	98.2	97.8	1.9
Injection	97.6	97.2	98.0	97.6	2.1
Average	98.4	98.1	98.6	98.3	1.5

It is clear from the results of Table 1 that TADLF is very accurate and stable for all the different types of attacks. The average detection accuracy for normal traffic, and also for more subtle types of attacks such as injection and spoofing (both >97.5%) is very high, 99.2%. The precision and recall values are all above 97%, which confirms the reliability in classifying malicious devices correctly, with very few false positives for malicious detection. The values for the F1-scores, which is close to the precision and recall, indicate stable behavior and performance particularly for the difficult types of attacks such as FDI and data modification. It is also noticeable that the average FAR is very low at 1.5% confirming the usefulness of the trust-aware system.

To provide a further validation of the developed TADLF framework, its results are benchmarked against those of commonly used IoT security/anomaly detection algorithms such as Random Forest, XGBoost, and normal LSTM models [3], [4]. Table 2 summarizes the results in terms of Accuracy, Precision, Recall, F1-Score, and False Alarm Rate (FAR) which illustrate the effectiveness of using trust-related features in hybrid CNN-BiLSTM model.

Table 3 Comparison with Baseline Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	FAR (%)
Random Forest	92.5	91.8	92.0	91.9	7.5
XGBoost	94.1	93.5	93.8	93.6	5.9
LSTM	96.3	96.0	96.2	96.1	3.7
TADLF (Proposed)	98.4	98.1	98.6	98.3	1.5

According to Table 2, the suggested TADLF outmatches the other baseline models based on every evaluation parameter. The conventional ML algorithms like Random Forest and XGBoost show a decent detection accuracy of 92.5% and 94.1% respectively, but have high FAR which restricts its feasibility for real-time IoT settings. The normal LSTM network learns the temporal relations among the data and improves detection, acquiring an accuracy of 96.3% and FAR of 3.7%. TADLF beats every other model with a 98.4% accuracy and a FAR of 1.5%.

Conclusion

In this work, a Trust-Aware Deep Learning Framework (TADLF) for detecting data integrity attacks, including false data injection attacks (FDI) in the IoT environment was proposed. By incorporating dynamic trust evaluation into a hybrid CNN-BiLSTM architecture, the proposed framework is well-suited for capturing both spatial correlation and temporal dependency within heterogeneous IoT data streams. In the system, a node-level trust is computed according to the consistency, stability, and historical integrity of data, which can distinguish between malicious nodes and noisy normal nodes. The performance of TADLF is evaluated on ToN-IoT dataset and shows much better detection performance over various machine learning and deep learning approaches

(e. G., LR, KNN, SVM, RF, XGBoost, CNN, LSTM). The proposed framework maintains a high accuracy(98.4%), precision (98.1%), recall (98.6%) and F1-score (98.3%), with an average false alarm rate of 1.5% for FDI, data integrity and injection attacks. A detailed comparison between our approach and other methods shows the benefits of incorporating trust-aware intelligence into hybrid CNN-BiLSTM modeling as well as its effectiveness over individual machine learning and deep learning approaches.

References

1. R. Huang, Y. Li, and X. Wang, "Attention-aware deep reinforcement learning for detecting false data injection attacks in smart grids," Int. J. Electr. Power Energy Syst., vol. 147, Art. no. 108815, May 2023.
2. M. N. Kurt et al., "Online cyber-attack detection in smart grid: A reinforcement learning approach," IEEE Trans. Smart Grid, vol. 10, no. 5, pp. 5174–5185, 2019.
3. C. Chen et al., "Data-driven resilient automatic generation control against false data injection attacks," IEEE Trans. Ind. Informat., vol. 17, no. 12, pp. 8092–8101, Dec. 2021.
4. V. P. Srinivasan et al., "Multi-label deep learning classification approach for false data injection attacks in smart grid," KSII Trans. Internet Inf. Syst., vol. 15, no. 6, pp. 2168–2187, 2021.
5. D. H. Hussein et al., "Integration of deep learning applications and IoT for smart healthcare," Indonesian J. Comput. Sci., vol. 14, no. 1, 2025.
6. N. Naik et al., "Hybrid deep learning-enabled framework for enhancing security and data integrity in healthcare IoT," Sci. Rep., vol. 15, Art. no. 31039, 2025.
7. S. A. Algethami and S. S. Alshamrani, "Deep learning-based cybersecurity framework for IoHT environments," Appl. Sci., vol. 14, no. 11, Art. no. 4729, 2024.
8. R. Yadav and A. K. Pradhan, "Wavelet probability distribution mapping for detection and correction of dynamic data injection attacks in WAMS," Int. J. Electr. Power Energy Syst., vol. 134, Art. no. 107447, 2022.
9. M. Yang and J. Zhang, "Data anomaly detection in IoT: Trends and challenges," Int. J. Adv. Comput. Sci. Appl., 2023.
10. L. Xiao et al., "Reinforcement learning-based false data injection attacks in smart grids," IEEE Trans. Ind. Informat., vol. 21, no. 4, pp. 3475–3484, Apr. 2025.
11. E. Shereen, K. Kazari, and G. Dán, "A reinforcement learning approach to undetectable attacks against automatic generation control," IEEE Trans. Smart Grid, vol. 15, no. 1, pp. 959–972, Jan. 2024.
12. A. Dinesh, K. Chakravarthi, A. K. Mulla, and P. Bhui, "Robust guaranteed cost output feedback control for real-time congestion management," IEEE Trans. Power Syst., vol. 39, no. 2, pp. 3349–3360, Mar. 2024.
13. D. T. Peng, J. Dong, J. Yang, and Q. Peng, "Dynamical failures driven by false load injection attacks against smart grid," IEEE Trans. Inf. Forensics Secur., vol. 17, pp. 2213–2226, Jun. 2022.
14. Y. Liu, P. Ning, M. K. Reiter, et al., "False data injection attacks against state estimation in electric power grids," ACM Transactions on Information and System Security, vol. 14, no. 1, pp. 1–33, 2011.
15. M. Ganjkhani, et al., "A novel detection algorithm to identify false data injection attack on power system state estimation," Energies, vol. 12, no. 11, p. 2209, 2019.