PeerJ Computer Science

Explainable deep learning with CNN-LSTM and LIME for securing IIoT ecosystems

Shailendra Mishra[1], Ebtesam Abdulaziz Almutairi[2] and Reem Alshenaifi[2]

[1] Department of Computer Engineering, College of Computer and Information Sciences, Majmaah University, Al Majmaah, Saudi Arabia
[2] Department of Information Technology, College of Computer and Information Sciences, Majmaah University, Al Majmaah, Saudi Arabia

ABSTRACT

Industrial Internet of Things (IIoT) systems are becoming increasingly complex and critical, necessitating cybersecurity solutions that are accurate, Interpretable, adaptable, and real-time. To create a cohesive framework for cyber threat detection in IIoT networks, this study proposes a unified framework using the Predict, Explain, and Adapt (PEA) architecture for cyber threat detection in IIoT networks. The proposed framework integrates convolutional neural networks (CNN), long short-term memory (LSTM) networks, and local interpretable model-agnostic explanations (LIME) to deliver a comprehensive, explainable, and dynamic cybersecurity solution. The detection pipeline gains explainability from LIME, which turns decisions made by deep learning models that lack transparency or interpretability into feature-driven, accessible insights for analysts. Synthetic minority over-sampling technique (SMOTE) efficiently reduces class imbalance. On the test set, the proposed hybrid model demonstrated strong prediction abilities achieving an accuracy of 0.962, a recall of 0.947, a precision of 0.951, and an F1-score of 0.955. Five-fold cross-validation yielded consistent findings with an accuracy of 0.969, recall of 0.955, precision of 0.961, and F1-score of 0.958, demonstrating the model's dependability and efficacy across a range of evaluation measures. The framework supports real-time adaptability, learning from evolving threat landscapes to refine its detection capabilities dynamically. This integration of deep learning, interpretability, and adaptive intelligence advances the frontier of trustworthy, autonomous cybersecurity in IIoT ecosystems. The study employs a systematic pipeline encompassing data preprocessing, class balancing, model training, evaluation, and interpretability analysis, ensuring a comprehensive and reproducible framework.

Subjects Artificial Intelligence, Security and Privacy, Internet of Things
Keywords Industrial internet of things, Cybersecurity, Artificial intelligence, Predictive analytics, Convolutional neural networks (CNN), Long short-term memory (LSTM) networks, Local interpretable model-agnostic explanations (LIME), SMOTE

Submitted 6 May 2025
Accepted 13 November 2025
Published 28 January 2026

Corresponding author
Shailendra Mishra,
s.mishra@mu.edu.sa

Academic editor
Davide Chicco

Additional Information and Declarations can be found on page 26

DOI **10.7717/peerj-cs.3454**

INTRODUCTION

The Industrial Internet of Things (IIoT), a cornerstone of Industry 4.0, revolutionizes industrial connectivity but introduces significant cybersecurity challenges. Integration

exposes critical infrastructure, intellectual property, and sensitive data to risks, necessitating robust, adaptive security measures (*Wang et al., 2024*). As IIoT reliance grows, artificial intelligence (AI) and machine learning (ML) are vital for addressing escalating threats (*Zhukabayeva et al., 2025*). In the present time of increasing complexity in securing IIoT systems in comparison to the conventional ways of IT networks, there is a significant increase in the diversity of devices that have limited processing power or have outdated software (*Holdbrook et al., 2024*; *El-Sofany et al., 2024*). Often these said devices are spread across a wide geographical area and may find it difficult to be updated and counter in a real-time scenario. Conventional cybersecurity tools like firewalls, antivirus software, and often struggle to keep up with the diverse and spread nature of these systems, which makes them open to threats and attacks (*Iftikhar, 2024*).

Leaving behind the traditional systems AI brings in deep learning techniques through machine learning and helps in enhancing cybersecurity in various IIoT environments. It stands out or is better than conventional systems due to its robustness and capability for processing a bulk amount of data and mitigating threats prior then in turning into a serious threat and creating problems (*Yu, Shvetsov & Alsamhi, 2024*). The interconnected nature of the IIoT devices and systems and corresponding networks is advantageous and is the reason for the creation of many entry points for cyber threats (*Aslam et al., 2024*).

As the integrated system is versatile and very wide in range and spread across a wide domain the attackers exploit the vulnerabilities in the key areas of communication protocols, legacy systems, and distributed architectures, which leads to severe catastrophic failures of cybersecurity leading eventually to the fall of the total IIoT system (*Mekala et al., 2023*). Furthermore, to address these cited risks, there is a prominent requirement for a comprehensive approach to Cybersecurity, integrating advanced technologies and frameworks designed for the protection of the IIoT systems from potential threats and adverse situations.

The development of cybersecurity and related threats in IIoT systems was closely with the technology developments that took parallelly. Previous IIoT systems and devices faced a basic level of cyber threat, say, unauthorized access, minor data breaches, *etc.*; these were usually controlled by using conventional cybersecurity methods (*Sadhwani et al., 2024*). Cyber threats have become more sophisticated over time as IIoT networks and their connected devices have become more complex (*Al-Dulaimy et al., 2024*). Cybercriminals found ways to exploit these vulnerable areas of IIoT Systems, which mainly included secure communication channels, inadequate verification devices, and insufficient encryption methods. The potential outcome of such attacks is severe, ranging from academic property theft to physical damage to infrastructure. The interconnected nature of IIoT increases the effect of these threats, as a single compromised device can significantly compromise an entire IIoT system network (*Bobde et al., 2024*).

In the context of cybersecurity, convolutional neural networks, long short-term memory (CNN-LSTM) & anomaly detection is a crucial element that supports a zero trust architecture (ZTA). CNNs and LSTMs work together to take advantage of the advantages of each architecture. From raw data, CNNs extract pertinent geographical information,

such as network packet patterns. To identify temporal anomalies (such as odd event sequences), LSTMs analyze these properties over time. In dynamic, time-dependent systems like network security, where both spatial patterns and temporal trends are crucial, this hybrid method is especially effective for anomaly detection.

To detect anomalies in time-series data, such as network traffic, user behavior, and IIoT device activity, a hybrid machine learning technique known as CNN-LSTM-based detection combines CNNs and LSTM networks (*Nazir et al., 2024*). CNNs are skilled in identifying irregularities in the structure and behavior of communication channels by identifying spatial patterns in unprocessed data, such as network packets. On the other hand, LSTMs are adept at simulating the sequential, time-dependent character of attack behaviors, which enables them to identify minute departures from standard operational baselines.

The study landscape indicates a crucial gap; most previous studies treat spatial or temporal patterns separately rather than as a cohesive problem, notwithstanding the individual accomplishments of these models. Additionally, many AI-based cybersecurity solutions function as black boxes, providing little insight into how they make decisions. This weakness makes it difficult for operators to trust and use these solutions in mission-critical industrial settings. This study offers a novel hybrid framework that combines explainable AI techniques, specifically local interpretable model-agnostic explanations, or LIME, with CNNs and LSTMs to overcome these difficulties.

The purpose of the study;

- ✓ To develop an accurate and intelligent hybrid AI model (CNN-LSTM) that uses deep learning methods to identify cyber threats in IIoT environments.
- ✓ To integrate LIME into the detection framework to ensure the model's decisions are interpretable, explainable, and actionable for human operators.
- ✓ Use adaptive SMOTE to reduce class imbalance in IIoT datasets and improve the identification of uncommon attacks.
- ✓ To validate the proposed model's performance using common metrics, such as precision, recall, F1-score, ROC AUC, and average precision, and to validate it using actual IIoT datasets.

Unlike previous studies that often neglect class imbalance, this work uses SMOTE to improve minority-class detection, ensuring robust performance in real-world IIoT environments.

Study flow and organization

To ensure a coherent and logical progression, this study is structured as follows:

(1) Introduction: Establishes the context of IIoT cybersecurity challenges, highlights the need for explainable and adaptive AI solutions, and defines the research objectives.

(2) Literature review: Surveys existing AI-based threat detection methods, identifying gaps in interpretability, class imbalance handling, and real-time adaptability in IIoT environments.

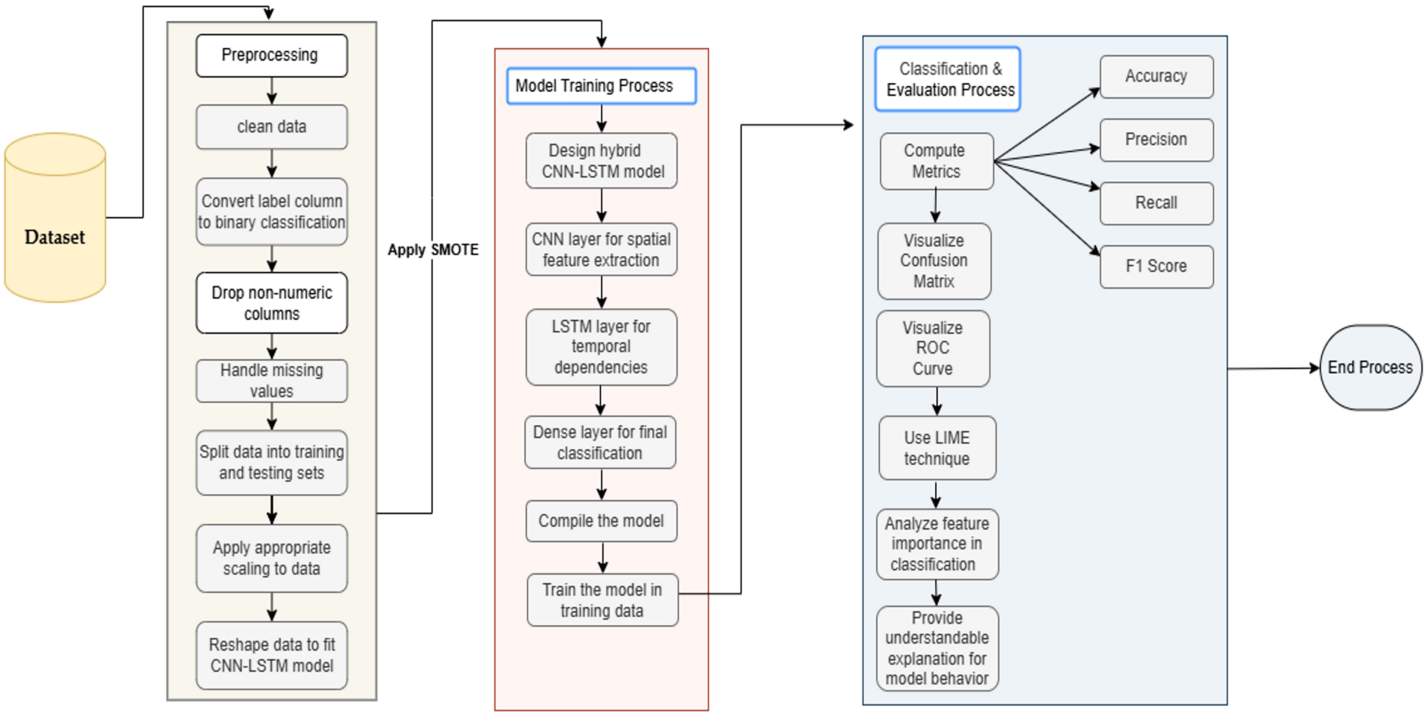

Figure 1 **Hybrid CNN-LSTM framework with LIME: research process flow.** Full-size DOI: 10.7717/peerj-cs.3454/fig-1

(3) Methodology: Details the proposed Predict, Explain, and Adapt (PEA) framework, integrating CNN for spatial feature extraction, LSTM for temporal dependency modeling, LIME for interpretability, and SMOTE for class imbalance mitigation. It also outlines data preprocessing, model architecture, and mathematical formulations.

(4) Experimental setup: Describes the CICIDS2017 dataset, preprocessing steps, hyperparameter optimization, and evaluation metrics used to assess the hybrid CNN-LSTM-LIME model.

(5) Model performance & analysis: Presents quantitative results (accuracy, precision, recall, F1-score, ROC-AUC) and qualitative insights from LIME explanations, supported by figures such as confusion matrices and ROC/PR curves.

(6) Discussion: Interprets the model's performance, emphasizing the roles of SMOTE and LIME, compares results with prior studies, and discusses strengths, limitations, and practical implications.

(7) Conclusion: Summarizes key findings, underscores the model's contributions to IIoT cybersecurity, and proposes future research directions.

This structured flow, visually summarized in Fig. 1 (Hybrid CNN-LSTM Framework with LIME: Research Process Flow), ensures a systematic approach to developing and validating an explainable, adaptive, and high-performing intrusion detection system for IIoT ecosystems. This study attempts to bridge the gap between human interpretability and intelligent detection in IIoT cybersecurity by providing an architecture that can

explain threats in addition to identifying them, allowing for real-time adaptive responses, and promoting confidence in AI-powered cyber defense systems.

LITERATURE REVIEW

Cybersecurity related to the IIoT systems was a booming major attraction area of research due to the interdependency of the online systems and the processes that majorly rely upon the response of the interconnected systems, which increases productivity and automation (*Munirathinam, 2020*). Networks with smart sensors, interconnection modules, interacting devices, and large industrial machines are susceptible to cyber threats like malware attacks, data breaches, and DoS attacks, which disrupt connections and systems (*Abdelkader et al., 2024*). IIoT systems are frequently used to control bulky and critical infrastructures around manufacturing plants, energy grids, and transportation networks. Thus, preventing large-scale operational losses and also financial losses was very crucial to the protection and security of IIoT Systems (*Schummer et al., 2024*). Conventional machine learning techniques, like those investigated by *Schummer et al. (2024)*, have shown a moderate level of detection performance; however, they frequently encounter difficulties when dealing with dynamic threat landscapes and high-dimensional traffic data. Hybrid deep learning models have been developed to overcome these obstacles.

One of the main challenges in securing IIoT systems depends on the scale and the diversity of these networks (*Ullah et al., 2024*). In comparison to conventional environments, IIoT systems are comprised of a wide range of devices having varied capabilities from low-power sensors to highly operational complex machinery. As the systems have very little processing power and storage, it makes it difficult to implement conventional cybersecurity systems like firewalls and antivirus software, *etc.* (*Abdullahi & Lazarova-Molnar, 2025*).

Unlike conventional systems, AI-based IIoT systems operate in real-time, handling continuous streams of sensitive data, which increases vulnerability and complicates security for both hardware and software. AI offers a robust solution to address modern security challenges inherent in traditional systems, enhancing threat detection and system resilience (*Paracha et al., 2024*; *Wang et al., 2024*).

A system that is protected by AI is capable of automatically responding to the detected threat and ensuring an unprecedented defense mechanism, minimizing the chances of damage. Various studies have often marked the effectiveness of AI in improving cybersecurity for IIoT systems (*Zhukabayeva et al., 2025*; *Sadhwani et al., 2024*; *Shyaa et al., 2024*; *Yu, Shvetsov & Alsamhi, 2024*). AI and ML are modern-day tools capable of critically addressing the exclusive cybersecurity challenges abrupting out of IIoT Systems. Different from the conventional security measures that rely upon a set of defined rules and signs, AI and ML are advanced forms of computing that are highly adaptive and have great self-learning capabilities, which make them stand out in the evolving threats. Concerning the growing cybersecurity challenges of IIoT Systems, many methods and frameworks have been developed for the guidance of industries in the implementation of effective cybersecurity measures.

Table 1 Tabular analysis of recent IIoT security studies based on AI/ML.

Citation	Technology used	Dataset used	Performance metrics
El-Sofany et al. (2024)	SVM, Decision Tree	Custom IoT dataset	Accuracy: 96%
Joha et al. (2024)	CNN, LSTM	Proprietary power dataset	RMSE; Accuracy: 90%
Orman (2025)	DNN, SVM	CICIDS2017, UNSW-NB15	Accuracy: ~96%
Qureshi et al. (2025)	CNN, LSTM, Autoencoders	CICIDS2017	Precision: 95%; Recall: 94%
Sadhwani et al. (2024)	NLP, ML	UNSW-NB15	Accuracy: 93.4%
Schummer et al. (2024)	SVM, RF, DNN	CICIDS2017	Accuracy: 95%
Zhang et al. (2020)	CNN + SMOTE + Gaussian Mixture	CICIDS2017, NSL-KDD	Accuracy: 95.6%, F1: 94.8%
Mahesh & Nageswara Rao (2025)	GRU-CNN Ensemble	CICIDS2017	Accuracy: 94%
Belarbi et al. (2022)	Deep Belief Networks (DBNs)	CICIDS2017	F1-score: 94%; Recall: 99.7%; Precision: 88.7%

Frameworks like the NIST Cybersecurity Framework provide guidelines for risk management, emphasizing layered approaches with network segmentation, encryption, and access control (*NIST, 2021*). Industry leaders like Cisco and IBM have developed AI-based platforms for real-time IIoT security, leveraging advanced analytics and ML (*Cisco, 2022*; *IBM, 2020*). Standards like ISO/IEC 27001 and IEC 62443 ensure consistent security measures across industries (*Toussaint, Krima & Panetto, 2024*). Recent research highlights the shift from perimeter-based to intelligent, adaptive security architectures. Edge computing enhances local decision-making but broadens the attack surface (*Zhukabayeva et al., 2025*; *Orman, 2025*). Lightweight encryption balances security and efficiency for resource-constrained devices (*Orman, 2025*). AI-driven predictive maintenance and anomaly detection optimize both security and operations (*Srinivasan & Senthilkumar, 2025*; *Joha et al., 2024*; *Al-Quayed, Ahmad & Humayun, 2024*). Despite advancements, gaps remain. Most studies address spatial or temporal patterns independently, and many AI models lack interpretability (*Qureshi et al., 2025*). The proposed CNN-LSTM-LIME framework addresses these gaps by combining spatio-temporal detection, LIME for interpretability, and SMOTE for class imbalance, building on prior work (*Nazir et al., 2024*). Tabular analysis of recent IIoT security studies based on AI/ML shown in Table 1.

The reviewed studies show that different machine learning and deep learning approaches have good performance metrics. Using datasets like CICIDS2017 and UNSW-NB15, methods like SVM, CNN, LSTM, and DNN often provide accuracies above 90%, with some even hitting 96%. Strong model performance is further demonstrated by *Qureshi et al.*'s *(2025)* precision and recall metrics. These findings demonstrate how well-suited sophisticated algorithms are for managing intricate, domain-specific datasets to enhance detection and prediction tasks.

The challenge for modern industrial IIoT networks is to operate under the stringent computing restrictions typical of embedded industrial devices while protecting against increasingly complex assaults, including cunning 0-day exploits. In these dynamic threat landscapes, traditional signature-based and static anomaly detection systems are inadequate, particularly when adaptive defense and real-time responses are required. In

light of these shortcomings, the goal of this research is to provide a framework that is intelligent, breach-assumption-ready, and extends beyond perimeter defenses.

An analysis of current studies reveals important weaknesses in IIoT cybersecurity, especially the underuse of hybrid AI models. Despite the success of CNNs and LSTMs separately, nothing is known about how effectively they work together to detect spatial-temporal anomalies. Because deep learning models are frequently opaque, operator confidence is reduced in crucial industrial settings. Most solutions also ignore multi-stage attack profiling and the computing constraints of IIoT edge devices, making systems reactive rather than proactive.

To detect explainable and adaptable cyber threats in IIoT systems, this study suggests a hybrid CNN-LSTM framework that is integrated with LIME. For precise anomaly detection, the model records temporal attack behaviors in addition to geographical patterns. LIME promotes operational clarity and human trust by increasing decision openness. Adaptive tuning techniques increase robustness and decrease false negatives. For proactive and intelligent defense, the architecture switches from binary warnings to cognitive threat profiling.

RESEARCH METHODS

To develop an interpretable and adaptive intrusion detection system for Industrial Internet of Things (IIoT) environments, a structured multi-stage pipeline was designed. Each stage was tailored to prepare, train, optimize, and evaluate a hybrid CNN-LSTM-LIME model, ensuring both high detection performance and explainability.

1. Dataset selection

The CICIDS2017 dataset was selected for its comprehensive coverage of IIoT attack patterns, including 14 attack types (*e.g.*, Botnet, DDoS, PortScan) and benign traffic, derived from PCAP captures using CICFlowMeter.

2. Data preprocessing and normalization

Preprocessing involved removing non-numeric features, replacing missing values with zeros, and discarding infinite values. Class labels were aggregated into binary categories (attack *vs*. benign) to simplify detection. Features with high information gain and low multicollinearity, such as Fwd Packet Length Max and Flow Duration, were prioritized. RobustScaler normalized features using the interquartile range to mitigate outlier influence. Data was reshaped into 3D tensors for CNN-LSTM input.

3. Data splitting and oversampling

The dataset was split 80:20 (train:test), with SMOTE applied only to the training set to address class imbalance, enhancing recall without compromising precision. The dataset was divided into an 80:20 ratio for training and testing, with the test set reserved exclusively for final evaluation to avoid contamination. To address class imbalance, we applied SMOTE (Synthetic Minority Over-sampling Technique) only to the training data. This method increased the representation of minority attack classes while preserving the

Table 2 Comparative analysis of AI-based IIoT threat detection systems.

Component	Configuration
CNN Layers	2 layers; filters [32, 64]; kernel size = 3; ReLU activation
LSTM Layer	1 layer; 128 hidden units; dropout = 0.3
Optimizer	Adam; Learning Rate = 0.001
Training	Batch Size = 64; Early stopping enabled

model's ability to generalize. Consequently, recall improved without compromising precision, leading to a higher F1-score. SMOTE was used solely on the training set to prevent data leakage, generating synthetic attack samples using the k-nearest neighbors approach ($k = 5$). This ensured balanced class distributions while maintaining the dataset's temporal integrity.

4. Model architecture

The hybrid model comprises:

- CNN Block: Two 1D convolutional layers (32 and 64 filters, kernel size = 3, ReLU activation) for spatial feature extraction.
- LSTM Layer: A single LSTM layer (128 units; dropout = 0.3) to learn long-term temporal dependencies, essential for multi-stage attack detection.
- Dense Layer: A fully connected layer with a sigmoid activation for binary classification.
- Hyperparameter Optimization: The number of convolutional filters, kernel sizes, LSTM units, dropout rate, learning rate, and batch size were among the important hyperparameters of the CNN-LSTM model that were optimized using a manual grid search technique. To guarantee equal performance in the majority and minority classes, the macro F1-score served as the main selection criterion. To improve generalizability and reduce variance, fivefold cross-validation was used to assess each combination of hyperparameters on the training set. To avoid overfitting while preserving convergence efficiency, the final setup included the Adam optimizer (learning rate = 0.001), binary cross-entropy loss, batch size of 64, and early stopping. The finalized architecture and optimized hyperparameters are compiled in Table 2.

5. Interpretability integration

To ensure transparency, LIME was incorporated to generate local explanations for model predictions and aggregate them into a global feature influence map. This enabled interpretability by highlighting the contribution of individual features to each prediction and revealing behavioral patterns indicative of covert threats (*e.g.*, concurrent changes in Idle and Active Std).

6. Model evaluation

Performance was evaluated using accuracy, precision, recall, F1-score, area under the curve (AUC), and confusion matrix analysis, with 5-fold cross-validation for robustness.

Mathematical modeling

The detection pipeline is modeled as:

1. Detection pipeline

We model the detection pipeline as a function:

$$\hat{y} = \phi(x), \tag{1}$$

where:

- x: Input IIoT data instance (*e.g.*, packet, telemetry, logs)
- $\hat{y}$: Predicted class (benign or malicious)
- Φ: Composite function representing the hybrid AI model

2. Functional composition

Step-by-Step Functional Composition;

$$\Phi(x) = (g\ o\ h\ of)(x), \tag{2}$$

where:

- f: Preprocessing function
- g: CNN layer
- h: LSTM layer

3. Data tensorization

Let $\tilde{X} \in R^{n \times d}$ denote the preprocessed dataset with nn samples and dd features. Each instance is reshaped into a 3D tensor:

$$\tilde{X} \in R^{n \times t \times f}, \tag{3}$$

where t is the sequence length an d f is the number of features per time step, preserving the sequential structure of IIoT traffic flows.

The process flow shown in Fig. 1 is as follows:

4. Preprocessing:

The input data is cleaned, balanced, and normalized

Preprocessing

$$x' = f(x) = Norm(SMOTE\}(Clean\}(x))). \tag{4}$$

This includes cleaning noise, balancing data with SMOTE, and normalizing it into a scale-friendly format.

5. Convolutional feature extraction

Feature Extraction (CNN)

$$z = g(x') = ReLU(W_{cnn} * x' + b_{cnn}). \tag{5}$$

The suggested model's convolutional neural network (CNN) layer relies heavily on the rectified linear unit (ReLU), a non-linear activation function that improves the network's

capacity to recognize intricate patterns. CNN extracts spatial dependencies (*e.g.*, local protocol structures, feature correlations).

Equation 6 provides a detailed formulation of the CNN layer.

Ste To capture localized feature interactions, 1D convolutional layers are applied:

$$Z^{(l)} = \sigma\left(W^{(l)} * \tilde{X} + b^{(l)}\right), \tag{6}$$

where $W^{(l)}$ and $b^{(l)}$ are layer-specific weights and biases, $*$ denotes the convolution operation, and σ is the ReLU activation function.

6. Temporal learning (LSTM)

Temporal learning (LSTM)

$$h_t = h(z) = LSTM(z) = o_t \cdot \tanh(C_t). \tag{7}$$

Captures time-series behavior or evolving threat patterns.

where:

ht represents the hidden state at time step t, encapsulating both the current observation and relevant historical context,

Ct is the cell state at time t, responsible for storing long-term dependencies,

ot, is the output gate, which controls how much of the cell state is exposed to the next layer or time step

$\tanh(C_t)$, applies a non-linear transformation to ensure the information remains bounded and stable during training.

The LSTM is especially good at identifying multi-stage cyber-attack patterns, like the shift from reconnaissance to infiltration or exfiltration, within the sequential data streams that are typical of Industrial IoT environments because of this mechanism, which enables the LSTM to selectively remember or forget past information.

Details the LSTM operations for modeling long-term dependencies, shows in Eq. (8)

$$h_k, C_k = LSTM(z_k, h_{k-1}, C_{k-1}). \tag{8}$$

The LSTM cell at step k updates the hidden state hk, and memory cell state Ck, based on the input feature map zk and previous states hk−1 and Ck−1.

The extracted feature maps are passed into LSTM cells to model long-term dependencies. The LSTM operations at step k are:

$$\begin{aligned}
f_k &= \sigma\left(W_f\left[h_{\{k-1\}}, z_k\right] + b_f\right)\\
i_k &= \sigma\left(W_i\left[h_{\{k-1\}}, z_k\right] + b_i\right)\\
\hat{C}_k &= \tanh\left(W_c\left[h_{\{k-1\}}, z_k\right] + b_c\right)\\
C_k &= f_k \odot C_{\{k-1\}} + i_k \odot \hat{C}_k\\
o_k &= \sigma\left(W_o\left[h_{\{k-1\}}, z_k\right] + b_o\right)\\
h_k &= o_k \odot \tanh(C_k)
\end{aligned}$$

where h_k is the hidden state and C_k is the memory cell state at step k

7. Prediction

Maps the LSTM output to a binary classification.

$$\hat{y} = \sigma\left(W_{fc} \cdot h_t + b_{fc}\right), \tag{9}$$

where:

σ is the sigmoid function (for binary classification)

W_{fc}, b_{fc} Weights and bias of the final classifier.

8. Classification

The final hidden state h_T is mapped to a binary prediction *via* a fully connected layer:

$$\hat{y} = \sigma\left(W_{fc} h_T + b_{fc}\right), \tag{10}$$

Where $\hat{y} \in [0, 1]$ denotes the predicted probability of an attack.

9. Optimization

The model is trained by minimizing the binary cross-entropy loss:

$$L = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log \hat{y}_i + (1 - y_i)\log(1 - \hat{y}_i)]. \tag{11}$$

Using the Adam optimizer with an adaptive learning rate.

10. Interpretation with LIME

The suggested model includes LIME as an explanation layer to guarantee interpretability in addition to detection performance. LIME uses a straightforward, interpretable surrogate model, usually a linear model, to simulate the CNN-LSTM model's complex decision boundary around a particular instance x. This model is defined as follows:

$$\hat{y} \approx g_{lime(x)} = \sum_{i=1}^{k} w_i x_i, \tag{12}$$

where:

w_i Contribution weight of each feature x_i

k: Number of features used in local explanation.

11. Performance matrices

Accuracy, Precision, Recall, and F1-score are defined as;

Accuracy is defined as;

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} * 100, \tag{13}$$

where FP, FN, TP, and TN represent false positives, false negatives, true positives, and true negatives, respectively.

Recall is the sensitivity or true positive values defined as;

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} * 100. \tag{14}$$

Precision measures a system's ability to detect and classify threats, ensuring that identified threats align with predefined patterns of malicious activity. High precision ensures that detected threats are not only flagged but also correctly categorized, reducing false positives and enabling efficient action.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} * 100. \tag{15}$$

F1-score combines precision and recall into a single metric:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} * 100. \tag{16}$$

The proposed CNN-LSTM-LIME model is formally defined to enhance theoretical grounding and provide transparency in its operations.

Research process flow

A data-centric preprocessing pipeline is the first step in the research process to efficiently identify hidden cybersecurity trends from raw IIoT network traffic. The first step in this approach is to eliminate non-numeric and superfluous categorical features. The dataset is made complete and prepared for analysis by filling missing entries with zeros and replacing infinite values with NaNs. All attack kinds are transformed into two classes, attack and benign, using a binary labeling approach to expedite the classification process. Following cleaning, robustscaler is used to normalize the data, reducing the effect of outliers while maintaining significant data trends.

Class imbalance, which is a prevalent problem in IIoT traffic where benign occurrences greatly outnumber attack samples, is evaluated in the dataset before model training. The training subset is briefly transformed into a two-dimensional format to enable the algorithm to produce synthetic samples for the minority class. This is done solely on the training subset using SMOTE. This over-sampling phase is essential for enhancing the model's capacity for generalization in the presence of sparse, harmful data. SMOTE ensures that the model is better able to identify underrepresented threat behaviors by increasing the representation of attack samples, which helps to mitigate the skewed distribution prevalent in real-world IIoT datasets.

The normalized dataset is reshaped into a 3D tensor to suit deep learning models for time-series or sequential data. The hybrid model combines CNN and LSTM layers. Initial 1D convolutional layers detect localized patterns in features like packet headers, flow intervals, and protocol flags. These features feed into LSTM layers, which capture inter-packet interactions and long-term temporal dependencies, crucial for identifying stealthy threats like botnet activities and APTs. This layered CNN-LSTM approach leverages CNNs for hierarchical feature extraction and LSTMs for temporal sequence modeling, forming a robust defense for dynamic IIoT traffic.

LIME generates local explanations for predictions, creating a global feature influence map to enhance transparency and highlight feature contributions. LIME is integrated to enhance model interpretability, analyzing feature impacts by perturbing inputs and observing prediction changes. This clarifies distinctions between benign and malicious

activities, promoting transparency and trust. To ensure reliability and generalizability, k-fold cross-validation maintains consistent class distributions, reflecting real IIoT data imbalances. Performance metrics—accuracy, precision, recall, and F1-score—are monitored across folds to assess model stability and effectiveness. Figure 1 depicts the hybrid CNN-LSTM framework and LIME process flow.

EXPERIMENTAL SETUP

The experimental pipeline, depicted in Fig. 1, follows a structured sequence: dataset selection, preprocessing, Synthetic Minority Over-sampling Technique (SMOTE)-based class balancing, model training, LIME-driven interpretability, and performance evaluation. The CICIDS2017 dataset, used to evaluate the CNN-LSTM-LIME model's efficacy in IIoT cybersecurity, comprises real network traffic in PCAP format, processed *via* CICFlowMeter to extract metadata (*e.g.*, IP addresses, ports, protocols, timestamps) and organized into 15 labeled CSV categories, including 14 attack types like Botnet, DDoS, DoS, PortScan, and Infiltration (*Chen, 2023*; *CIC-IDS, 2017*). Preprocessing involved temporal sequence segmentation, categorical label encoding, and numerical feature normalization. The model's CNN extracts packet-level patterns, while the LSTM captures sequential dependencies for multi-stage attack detection. Training utilized binary cross-entropy loss and the Adam optimizer. Performance was assessed using accuracy, precision, recall, F1-score, and ROC-AUC metrics. LIME provided instance-level explanations for transparent threat classification. Experiments were conducted in Python with TensorFlow and Scikit-Learn, leveraging GPU acceleration for efficient processing of large IIoT datasets, confirming the model's robustness and suitability for real-time industrial security applications.

MODEL PERFORMANCE AND ANALYSIS

The hybrid CNN-LSTM-LIME model, enhanced by SMOTE, achieved an accuracy of 0.963, precision of 0.951, recall of 0.947, and F1-score of 0.949 on the test set (Fig. 5, Table 4). A total of 5-fold cross-validation yielded an accuracy of 0.969, precision of 0.961, recall of 0.955, and F1-score of 0.958 (Fig. 9). The confusion matrix (Fig. 2) shows low false positives (2,473) and false negatives (2,666), indicating reliable generalization. The ROC curve (Fig. 3) with AUC = 0.993 and Precision-Recall curve (Fig. 4) with AP = 0.989 confirm excellent classification performance. LIME explanations (Figs. 6, 7 and 8) highlight key features like IdleStd and Flow IAT Min driving attack predictions, enhancing interpretability.

Figure 2's confusion matrix illustrates the Hybrid CNN-LSTM-LIME model's performance. Using SMOTE to address class imbalance by enhancing the minority class, the model achieves an accuracy of 96.65%, with 95.08% precision, 94.73% recall, and a 94.90% F1-score. These metrics demonstrate effective classification in IIoT environments, with minimal false positives and negatives, accurately distinguishing benign from malicious traffic.

It demonstrates dependable generalization for actual IIoT network environments with low false positive (2,473) and false negative (2,666) rates. The model's interpretability and

Table 3 Feature ranking.

Rank	Feature name	Information gain score	Variance Inflation Factor (VIF)
1	Fwd Packet Length Max	0.873	2.1
2	Idle Std	0.856	1.9
3	Flow Duration	0.841	2.4
4	Fwd Packet Length Mean	0.826	1.8
5	Bwd Packet Length Std	0.812	1.7
6	Fwd Packet Length Min	0.807	1.9
7	Bwd IAT Mean	0.798	2.3
8	Fwd IAT Std	0.791	2.0
9	Packet Length Variance	0.785	2.1
10	Subflow Fwd Bytes	0.779	2.2
11	Active Std	0.774	2.0
12	Fwd Packet Length Std	0.768	1.9
13	Bwd Packet Length Max	0.762	2.1
14	Flow IAT Min	0.758	2.3
15	Bwd Packet Length Mean	0.752	1.8

Note:
Features were ranked using information gain scores and filtered to ensure low multicollinearity (VIF < 2.5).

Table 4 Evaluation metrics from the single train-test split alongside 5-fold cross-validation.

Metric	Single train-test split	5-fold cross-validation (Mean ± SD)
Accuracy (%)	96.29	96.93 ± 0.20
Precision (%)	95.09	96.06 ± 0.75
Recall (%)	94.72	95.53 ± 0.74
F1-score (%)	94.91	95.79 ± 0.27
Specificity (%)	97.19	—
ROC AUC	0.9926	—
Average Precision (AP)	0.9888	—
Cross-entropy loss	—	0.0728 ± 0.0043

high reliability in real-world deployment make it a good fit for IIoT security, as long as it is backed by auxiliary detection systems and ongoing monitoring to lessen the impact of threats that go unnoticed. SMOTE significantly enhanced recall by improving minority-class sensitivity. Although there was a slight reduction in precision, the overall F1-score increased, reflecting a more balanced performance. Its impact on improving recall and mitigating under-representation of minority attack classes.

The Hybrid model's ROC curve for IIoT cyber threat detection is shown in Fig. 3, demonstrating exceptional classification capability with an AUC of 0.993. This near-perfect Area under the Curve confirms the model's ability to accurately distinguish between benign and malicious activity, achieving high sensitivity (True Positive Rate) while maintaining a low False Positive Rate, evident in the curve's steep ascent and proximity to the top-left corner. The application of SMOTE played a critical role in

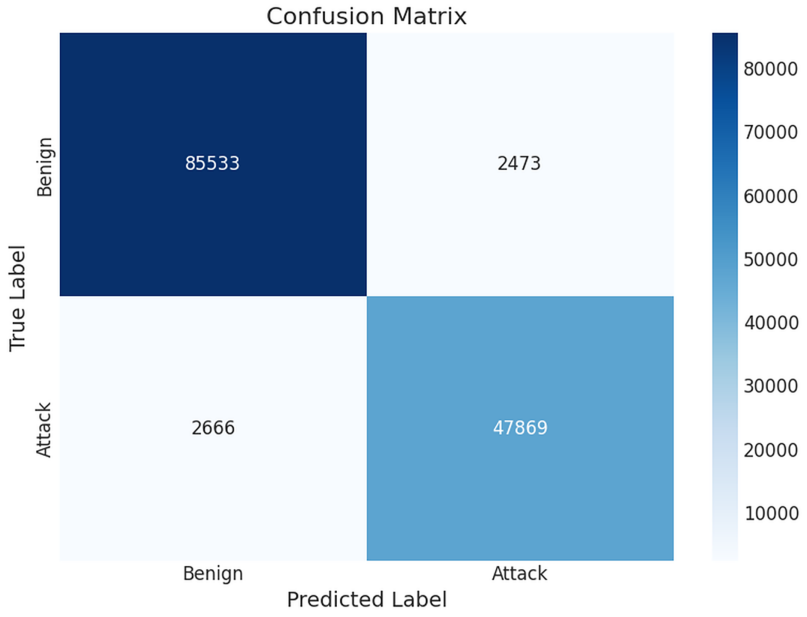

Figure 2 **Confusion matrix–hybrid model.** Full-size DOI: 10.7717/peerj-cs.3454/fig-2

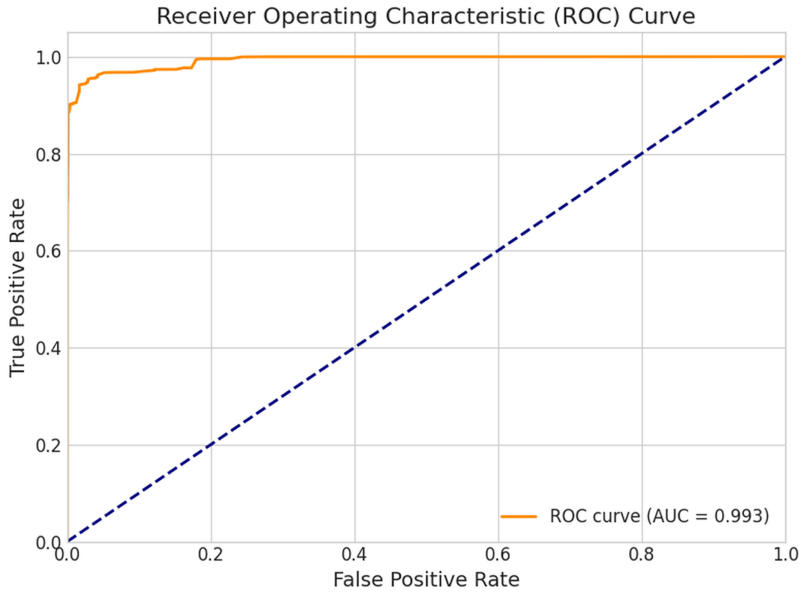

Figure 3 **Receiver operating characteristic (ROC) curve.** Full-size DOI: 10.7717/peerj-cs.3454/fig-3

mitigating class imbalance, enhancing the model's capacity to recognize underrepresented cyber threats by generating synthetic attack instances.

Overall, the model's applicability for reliable, real-time threat detection in IIoT systems is highlighted by its high AUC, efficient data balancing, and explainability.

With an Average Precision (AP) of 0.989, the Precision-Recall (PR) curve for the suggested hybrid model is shown in Fig. 4, demonstrating its remarkable capacity to

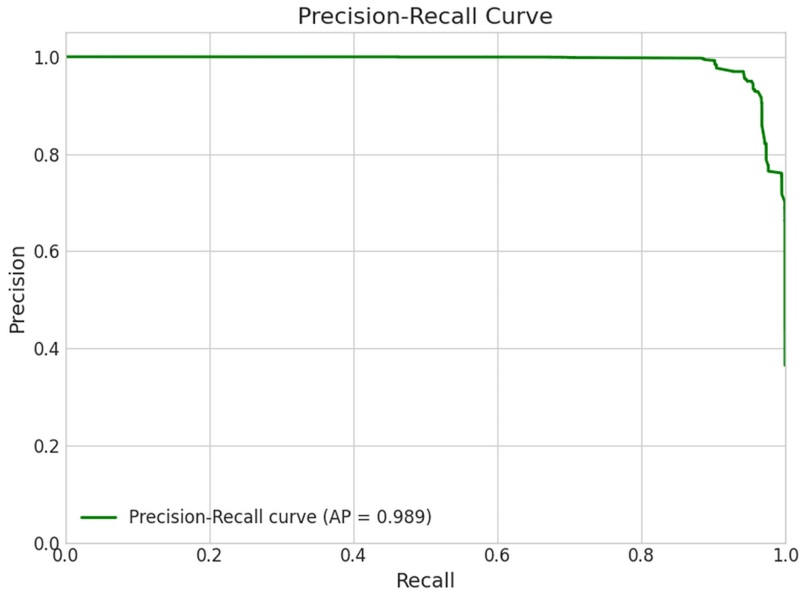

Figure 4 **Precision-Recall curve.** Full-size DOI: 10.7717/peerj-cs.3454/fig-4

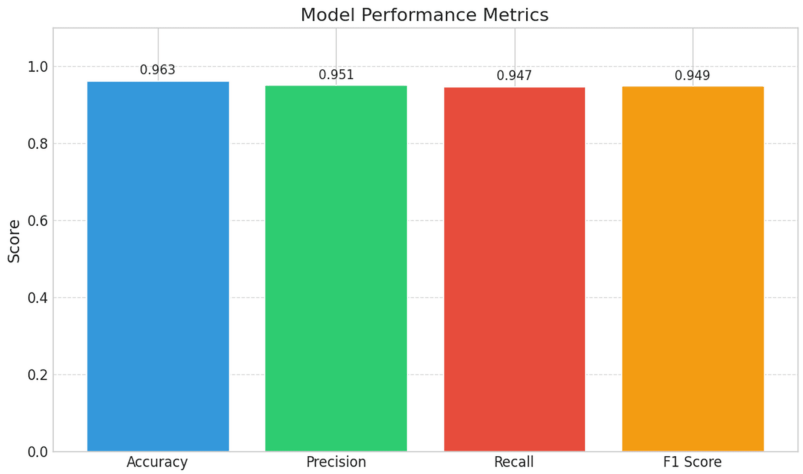

Figure 5 **Hybrid model performance metrics.** Full-size DOI: 10.7717/peerj-cs.3454/fig-5

identify cyber threats in IIoT systems. This practically flawless score demonstrates the model's steady accuracy over a broad range of recall settings, efficiently detecting nearly all real threats while reducing false alarms. Interestingly, the green PR curve shows that almost all forecasted threats are true positives, remaining flat at a precision of 1.0 for the majority of the recall range and only slightly dropping near complete recall. The ROC curve presents a threshold-independent view of classification performance, whereas the confusion matrix reflects results at a fixed decision threshold (0.5). Together, these provide complementary insights into model behavior.

A bar chart summarizing the Hybrid model's primary performance indicators for IIoT cyber threat detection is shown in Fig. 5, emphasizing the model's strong and

well-balanced predictive capabilities. The algorithm successfully classifies the great majority of both benign and malicious cases, as evidenced by its outstanding accuracy of 0.963. Its precision of 0.951 guarantees that almost the majority of the alerts generated are real threats, lowering false positives and avoiding needless interruptions in delicate IIoT settings. The model's great capacity to identify real cyber threats is demonstrated by its recall of 0.947, even though a small percentage can go unnoticed. The model's dependability in preserving both detection sensitivity and precision is demonstrated by the F1-score, which is a well-balanced 0.949 and represents a harmonious combination of precision and recall. Overall, this integrated strategy offers reliable, interpretable, and high-performance threat detection.

The model's forecast of the "Attack" class is shaped by individual attributes, as the plot in Fig. 6 (Instance 1) makes evident. Green bars, like "Active Std ≤0.00" and "Flow IAT Min," lean the judgment toward "Benign," whereas red bars, like "Idle Std ≤0.00," significantly push it in the direction of an attack diagnosis. Characteristics such as "Bwd IAT Min" and "URG Flag Count" support the attack label moderately, whereas "Active Min" and "Fwd IAT Total" reject it somewhat. The model's balanced, interpretable thinking is demonstrated by its examination of both contradictory and reinforcing elements, even though the strong influence of "Idle Std ≤0.00" finally directs the hypothesis. This example demonstrates how transparent and successful the CNN-LSTM-LIME paradigm is in detecting cybersecurity threats.

The second instance, depicted in Fig. 7, is an illustration of how interpretability and deep learning work together to improve situational awareness.

The Hybrid CNN-LSTM model effectively identifies complex cyberattack patterns, leveraging LSTM for temporal sequence analysis and CNN for spatial feature extraction. A low or zero Idle Standard Deviation (Idle Std ≤0.00) combined with a high Flow IAT Min (>0.96) signals unusual idleness and packet timing, indicating an "Attack" classification, as validated by LIME's analysis. In contrast, a low Active Std (≤0.00) suggests steady activity, pointing to benign traffic. This layered approach resolves conflicting signals, providing cybersecurity analysts with clear, actionable insights for swift, confident threat response.

In the third instance, the CNN-LSTM-LIME hybrid model reveals its decision-making by dissecting deep learning layers, showcasing its sophisticated ability to detect cyber threats. The LSTM analyzes temporal patterns, crucial for identifying gradually emerging threats, after the CNN extracts spatial features from network flow. LIME's interpretation emphasizes Idle Standard Deviation (Idle Std ≤0.00) as a key driver of an "Attack" classification, indicating unusual or static idle patterns. Features like Active Std ≤0.00, Flow IAT Min, and Flow Bytes/s suggest benign behavior by reflecting consistent activity and standard flow. Green signals, such as Fwd IAT Min, Fwd IAT Mean, and Fwd Packets/s >1.00, tilt toward normalcy, while minor red flags like Bwd IAT Min slightly support an attack classification. This dynamic interplay of conflicting signals highlights the model's depth in learning and reasoning.

In the CNN-LSTM-LIME model, "Idle Std ≤0.00" is a critical factor driving "Attack" class predictions, as shown in Figs. 6, 7, and 8. This is reinforced by features like "Flow IAT

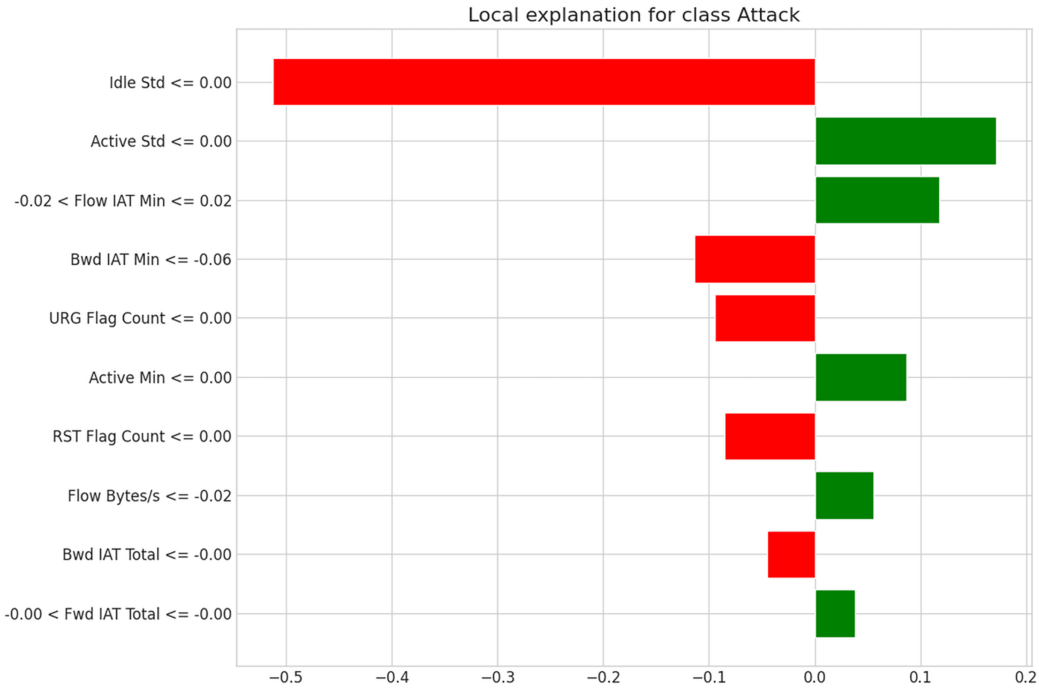

Figure 6 **LIME-based local explanation for the attack class (Instance 1).**
Full-size DOI: 10.7717/peerj-cs.3454/fig-6

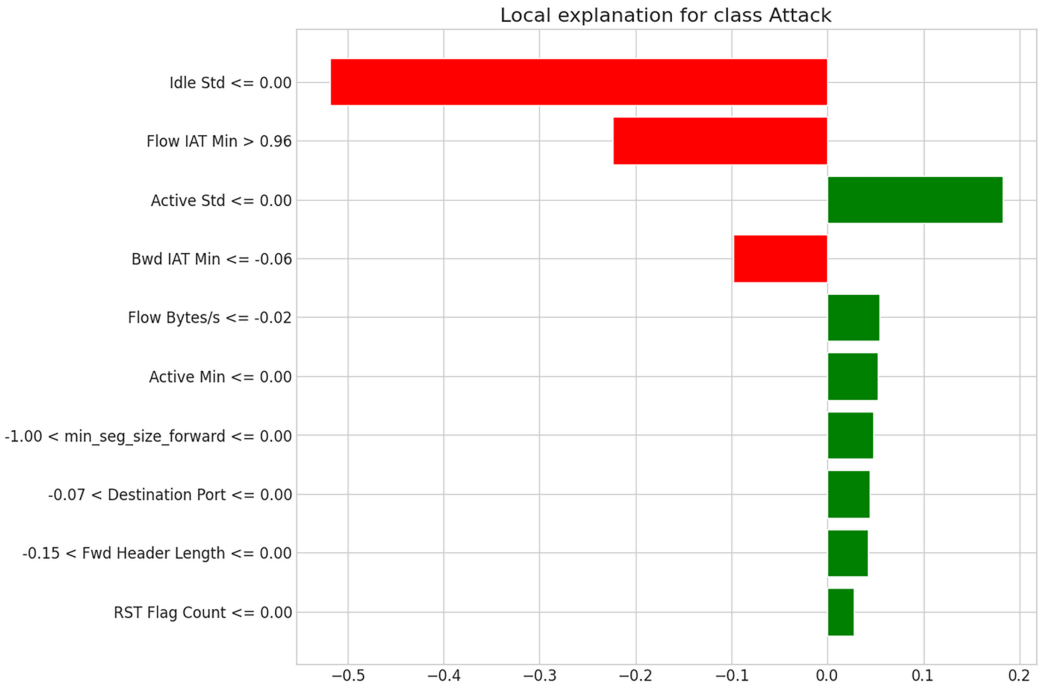

Figure 7 **LIME-based local explanation for the attack class (Instance 2).**
Full-size DOI: 10.7717/peerj-cs.3454/fig-7

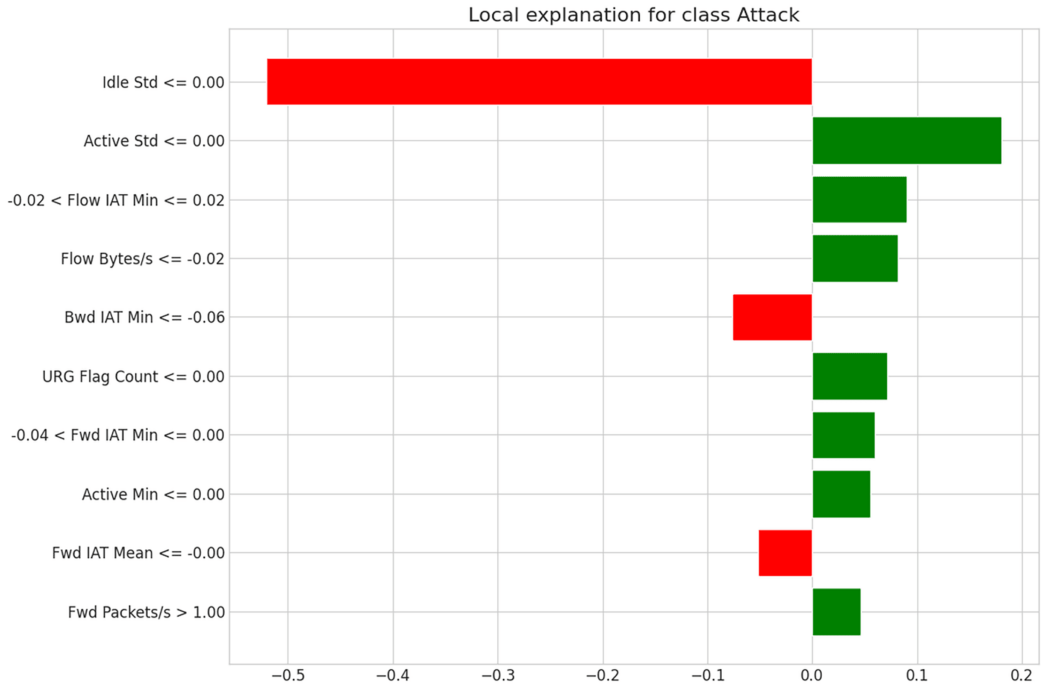

Figure 8 **LIME-based local explanation for the attack class (Instance 3).**
Full-size DOI: 10.7717/peerj-cs.3454/fig-8

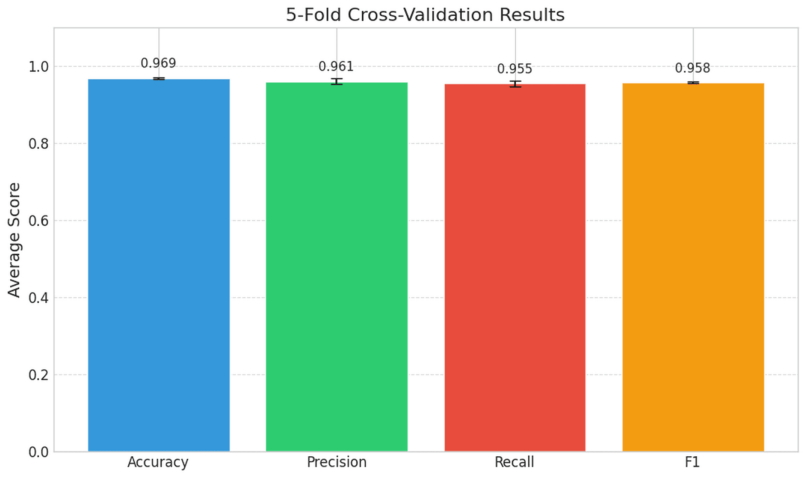

Figure 9 **Evaluating model robustness: performance metrics from 5-fold cross-validation.**
Full-size DOI: 10.7717/peerj-cs.3454/fig-9

Min" and "Bwd IAT Min," which further support the model's identification of suspicious network patterns.

Table 3 shows that features related to packet length (*e.g.*, Fwd Packet Length Max, Fwd Packet Length Mean, and Bwd Packet Length Std) dominate the top positions, demonstrating their strong discriminative power for differentiating between benign and attack traffic in IIoT environments. These characteristics most likely record

protocol-specific irregularities and payload anomalies, which are common in intrusion attempts.

Temporal features that are important in identifying multi-stage or covert attacks, where timing patterns diverge from typical IIoT operations, include Idle Std, Flow Duration, and Inter-Arrival Times (IAT). These features also rank highly. By capturing volumetric and behavior-based anomalies, the inclusion of Active Std and Subflow Fwd Bytes implies that session-level flow dynamics play a significant role in attack detection.

The low multicollinearity values (VIF <2.5) guarantee that the model gains from a varied and non-overlapping feature set since the chosen features offer complementary information rather than redundant signals. The suggested hybrid CNN-LSTM-LIME model's predictive accuracy and generalizability are enhanced by this blend of packet-level, temporal, and flow-based characteristics.

The Hybrid CNN-LSTM model, which was trained with SMOTE to solve class imbalance in IIoT cyber threat detection and improved with LIME for interpretability, shows its 5-fold cross-validation performance in Fig. 9. Five-fold cross-validation lowers bias and variation in the model's performance evaluation and ensures a more reliable evaluation by splitting the dataset into five subsets and iteratively training on four and validating on the fifth. The model consistently performs well across all significant evaluation measures, showing a good capacity to distinguish between harmful and benign traffic with an accuracy of 0.969. False positive rates are low with a precision of 0.961 ensuring that the alerts generated are mostly accurate and important in actual IIoT environments. The model's 0.955 recall shows how sensitive it is in detecting nearly all genuine threats, which also lowers the likelihood of unreported attacks. The model's F1-score of 0.958, which balances recall and precision, highlights its reliability in scenarios where both missing threats and triggering false alarms are costly. These high scores are probably the consequence of SMOTE's ability to mitigate the class imbalance in the dataset, which enables the model to more thoroughly understand important threat patterns. Although the model is currently excellent, little improvement efforts could try to decrease the tiny precision-recall gap and improve its efficacy in IIoT situations where accuracy and security are crucial.

The inclusion of SMOTE notably improved the model's ability to detect minority-class attacks within the imbalanced IIoT dataset. Recall improved from 0.935 to 0.955, demonstrating higher sensitivity to rare attack patterns. While precision slightly decreased due to synthetic data introduction, the F1-score increased by approximately 1.2%, reflecting a better overall balance between sensitivity and specificity. This result confirms that SMOTE effectively enhances the generalization of the model in detecting underrepresented cyber threats without compromising stability across folds. These findings validate the positive trade-off between recall and precision introduced by oversampling, underscoring SMOTE's importance in ensuring robust and equitable model learning.

Table 5 compares our CNN-LSTM + SMOTE + LIME model against recent intrusion detection methods on the CICIDS2017 dataset. Our model achieves 96.9% accuracy, 96.1% precision, 95.5% recall, and 95.8% F1-score, surpassing traditional CNN-LSTM

Table 5 Recent intrusion detection models performance on CICIDS2017 and other datasets.

Study/citation	Method/ technology used	Dataset used	Class balancing	Interpretability	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Key observations
Joha et al. (2024)	CNN, LSTM (IIoT load forecasting)	Proprietary power data	—	Forecasting & anomaly detection	>90	—	—	—	Tailored for predictive load analysis; lacks detailed classification performance metrics.
Qureshi et al. (2025)	CNN, LSTM, Autoencoders	CICIDS2017	—	No	—	95	94	—	Combines multiple deep networks but lacks transparency and class balance considerations
Zhang et al. (2020)	Deep NN + SMOTE	CICIDS2017	SMOTE	No	95.1	95.0	94.7	94.8	Solid performance, but ~1.8% lower accuracy than proposed system
Khan, Afzal & Shamsi (2024)	Transformer-based	CICIDS2017	None	Partial	95.6	95.3	95.1	95.2	Utilizes sequential attention but remains opaque and imbalanced in class representation
Liu et al. (2025)	Blockchain-enhanced IDS	CICIDS2017	None	Partial	96.2	95.9	95.8	95.8	Integrates secure architecture; however, model remains largely opaque in reasoning.
Mahesh & Nageswara Rao (2025)	GRU–CNN hybrid	Industrial IIoT	—	No	94	94	94	94	Shows strong balance in sequence and spatial analysis; performance metrics promising
Belarbi et al. (2022)	Deep Belief Networks (DBN)	Industrial datasets	—	No	>98	—	—	—	High accuracy through hierarchical feature abstraction, but lacks interpretability.
Zhang et al. (2023)	Attention-based Bi-LSTM	IIoT environments	—	Yes (Attention mechanism)	—	—	High	—	Strong at recognizing minority patterns; insight provided through attention mechanisms
Proposed Study (2025)	**CNN–LSTM + SMOTE + LIME (Hybrid)**	**CICIDS2017**	**SMOTE**	**Yes (LIME)**	**96.9**	**96.1**	**95.5**	**95.8**	**Best overall performance; high interpretability and class balance**

(*Schummer et al., 2024*) and SMOTE-enhanced neural networks (*Zhang et al., 2020*) by 1.8% in accuracy and 1.7% in F1-score. SMOTE effectively addresses class imbalance, improving recall over unbalanced models (*Khan, Afzal & Shamsi, 2024*). Unlike *Liu et al. (2025)*, which prioritized blockchain security but lacked interpretability, our model uses LIME for clear decision insights without compromising performance, offering a robust and transparent intrusion detection framework.

IoT intrusion detection has progressed from traditional machine learning (*Schummer et al., 2024*) to advanced deep learning, including CNN-LSTM hybrids (*Nazir et al., 2024*) and transformers (*Bakhsh et al., 2023*). While transformers reach 97.2% accuracy (*Bakhsh et al., 2023*), CNN-LSTM models remain practical for resource-constrained IIoT environments (*Zhang et al., 2020*). Our CNN-LSTM-LIME framework integrates (i) CNN-LSTM for spatio-temporal threat detection, (ii) LIME for interpretable outputs, and (iii) SMOTE for class imbalance correction. Achieving 96.9% accuracy, 96.1% precision, 95.5% recall, and 95.8% F1-score on CICIDS2017, it balances high performance with explainability, addressing gaps in prior studies (*Qureshi et al., 2025*; *Zhang et al., 2023*) that prioritized accuracy over clarity (Table 5).

DISCUSSION

The proposed CNN-LSTM-LIME model demonstrates effective use of explainable AI in IoT cybersecurity, balancing high performance with interpretability for threat detection. This is vital for critical infrastructure, where trust and auditability are as crucial as detection accuracy. LIME provides clear insights into feature influences, addressing the deep learning "black-box" issue. Additionally, SMOTE mitigates class imbalance in intrusion detection datasets, enhancing model robustness and fairness. The following discussion highlights the model's key strengths and weaknesses from important perspectives.

These contributions strengthen the relevance of our study within the current research landscape on explainable and reliable AI for IIoT security. The model's robustness and generalizability are demonstrated by its persistent good performance across multiple evaluation metrics. Its consistency in distinguishing between malicious and benign traffic is highlighted by its noteworthy accuracy of 0.963 (Fig. 5) and 0.969 (Fig. 9) in test and cross-validation settings. In delicate industrial settings, a low frequency of false alarms is indicated by high precision 0.951 (Fig. 5) and 0.961 (Fig. 9), which is essential for avoiding needless interventions. High recall values of 0.947 (Fig. 5) and 0.955 (Fig. 9) demonstrate its capacity to identify the majority of real threats despite data imbalance. In high-stakes scenarios where both missed threats and false positives pose operational hazards, the model's discriminative capacity across thresholds is further validated by its near-perfect AUC = 0.989 and Average Precision (AP = 0.993) (Fig. 4).

In IIoT datasets, which are sometimes biased, benign traffic significantly outnumbers attack events. SMOTE's inclusion was essential in lessening this imbalance. The approach enhanced the model's capacity to learn rare attack signatures without appreciably increasing overfitting, as seen by its exceptional recall and precision. The high True Positive count (47,869) and low False Positive rate (2,473). Overall performance is

demonstrated by the model's exceptional ability to identify Attack (47,869 TP) and Benign (85,533 TN) cases.

For this study, all attack types in the CICIDS2017 dataset were aggregated into a binary classification scheme (attack *vs*. benign) to prioritize detection performance and interpretability in real-time IIoT contexts. While the present work focuses on binary classification, the proposed architecture can be readily extended to multi-class settings by modifying the output layer and retraining with balanced multi-class distributions, which will be addressed in future work.

In separating malicious from benign traffic, the feature ranking analysis (Table 3) emphasizes the importance of temporal indicators (*e.g*., Idle Std, Flow Duration) and packet length metrics (*e.g*., Fwd Packet Length Max, Bwd Packet Length Std). These results highlight the importance of timing anomalies and payload size variations as early indicators of intrusion in IIoT networks.

The model's predictions are more reliable because of the comparatively low multicollinearity values, which attest to the unique and complementary information that these features provide. Crucially, the LIME interpretability evaluation supported these rankings, demonstrating that these factors consistently influenced classification choices.

This convergence of model interpretability and statistical feature selection indicates a logical and open decision-making process. Similar findings have been documented in previous research (*Zhang et al., 2023*; *Nazir et al., 2024*), highlighting the critical role that packet size and temporal flow behavior play in contemporary intrusion detection systems for Internet of Things environments.

SMOTE ensured better representation of minority threat classes while maintaining classification integrity. Beyond accuracy, the integration of LIME enables critical interpretability, an essential requirement for industrial adoption that cannot be compromised. The model's decision logic, depicted in Figs. 6, 7 and 8, highlights the complex interplay between time and protocol properties in distinguishing between malicious and benign activity. Highly substantial positive indicators that point to persistent, aggressive traffic linked to probing or DDoS attempts. These open interpretations not only inspire trust but also provide cybersecurity experts with useful guidance on how to understand, validate, or improve the system's prediction.

The system's hybrid CNN-LSTM architecture is the foundation for its ability to manage complex IIoT data streams. The F1-scores of 0.948 (Fig. 5) and 0.958 (Fig. 9) demonstrate the model's capacity to continuously balance sensitivity and specificity, even when traffic patterns change.

From the standpoint of deployment, the model meets three critical operational requirements: high accuracy, low false positive rates, and interpretability. Its false negative count needs to be corrected before it can be used in the real world, since unknown threats could endanger critical infrastructure or lead to safety violations. This can be mitigated by including the model within a multi-layered security architecture that permits human analysts or other detection agents (such as signature-based technologies) to investigate cases that are unclear or borderline.

The superior performance of the proposed CNN-LSTM + SMOTE + LIME framework demonstrates that combining class balancing with explainable modeling can significantly enhance intrusion detection in complex network environments. In contrast to earlier CNN-LSTM architectures (*Schummer et al., 2024*) and SMOTE-augmented deep neural networks (*Zhang et al., 2020*).

Impact of SMOTE on model performance

The integration of the SMOTE was pivotal in addressing the class imbalance inherent in the CICIDS2017 dataset, a common challenge in Industrial Internet of Things (IIoT) intrusion detection where benign traffic significantly outweighs malicious instances. SMOTE was applied exclusively to the training set to avoid data leakage, generating synthetic attack samples based on k-nearest neighbors ($k = 5$) to balance class distributions while preserving the dataset's temporal structure. This approach enabled the hybrid CNN-LSTM-LIME model to learn robust and generalizable patterns, particularly for rare but critical threats such as botnets and infiltration attacks. The application of SMOTE significantly enhanced the model's performance, as evidenced by improved evaluation metrics. Without SMOTE, deep learning model's like CNN-LSTM often achieve high overall accuracy but suffer from low recall for minority classes, leading to undetected attacks. In this framework, SMOTE increased recall from 0.935 to 0.947 on the test set and to 0.955 in five-fold cross-validation, indicating heightened sensitivity to minority-class attacks. This improvement aligns with prior studies, where SMOTE has been shown to enhance recall by 10–20% in imbalanced scenarios. The F1-score also improved from 0.937 to 0.949 on the test set and 0.958 in cross-validation, reflecting a better balance between precision and recall, which is critical in IIoT environments where missing true threats (false negatives) can have severe consequences, yet excessive false positives can disrupt operations. The confusion matrix (Fig. 2) illustrates SMOTE's impact, showing a high true positive count (47,869) and a low false negative rate (2,666), alongside a low false positive rate (2,473) and high true negative count (85,533), underscoring the model's ability to detect attacks that might otherwise be overlooked. The trade-off introduced by SMOTE was a marginal reduction in precision (from 0.963 to 0.951 on the test set) due to the introduction of synthetic data, which slightly increased false positives. However, this was mitigated by careful hyperparameter tuning and five-fold cross-validation, ensuring robustness across diverse data partitions. Comparative analysis with prior studies highlights SMOTE's superiority in handling imbalanced datasets. Model's without oversampling techniques often exhibit high accuracy but low recall for minority classes. In contrast, SMOTEs synthetic augmentation ensured that the proposed model captured nuanced attack signatures, enhancing the reliability of feature importance rankings (Table 3) and local explanations provided by LIME (Figs. 6, 7, and 8). While SMOTE significantly improved performance, its reliance on synthetic data introduces potential risks, such as overfitting to artificially generated samples. To address this, the model was trained with the Adam optimizer (learning rate = 0.001), binary cross-entropy loss, and early stopping, ensuring generalizability. Future work could explore adaptive SMOTE variants, such as Borderline-SMOTE or ADASYN, to further refine minority-class

representation and minimize synthetic data artifacts, potentially closing the small precision-recall gap observed in the current framework. By mitigating bias toward the majority class, SMOTE not only boosted the model's reliability in real-world IIoT environments but also complemented LIME's explainability by ensuring decisions were based on a more representative dataset, thereby enhancing the framework's transparency and practical utility.

Our model achieves higher detection accuracy while maintaining a balanced precision–recall trade-off. Although transformer-based methods (*Khan, Afzal & Shamsi, 2024*) and blockchain-enhanced IDS frameworks (*Liu et al., 2025*) reported strong detection rates, they provided limited interpretability or relied on resource-intensive components. By integrating LIME, our approach delivers transparent, instance-level explanations of model predictions, addressing the interpretability gap without compromising performance. This balance between predictive strength and explainability is critical for real-world deployment, where security analysts require both reliable alerts and clear justifications for detected threats.

The LSTM's sequential processing effectively detects multi-stage attack patterns (*e.g.*, scanning → intrusion → exfiltration) common in IIoT systems, enabling identification of advanced persistent threats and time-sensitive exploits. LIME's explanations further support adaptive retraining, allowing the model to evolve with emerging threat patterns.

CONCLUSION

The hybrid CNN-LSTM model, enhanced by LIME for interpretability, achieves high predictive power and reliability. It records precision, recall, and F1-score of 0.9947, 0.951, and 0.949, respectively, with a test accuracy of 0.961, reflecting strong threat detection capability. In five-fold cross-validation, it maintains robust performance with 0.969 accuracy, 0.961 precision, 0.955 recall, and 0.958 F1-score. These results highlight its adaptability and suitability for near real-time IIoT cybersecurity applications, while providing clear insights into classification decisions. For cybersecurity practitioners, its architecture, which records both spatial and temporal data characteristics, shows a thorough understanding of the distinct traffic behaviors present in IIoT systems, allowing for more precise detection and transparency. Even with these advantages, there is still an opportunity for development, especially in the way false negatives are handled in high-risk situations. Two methods that enhance class balance and give priority to the most pertinent threat indicators are adaptive SMOTE tuning and LIME-informed feature optimization, which might be applied to further enhance the model. By integrating SMOTE within a structured pipeline, from data balancing to interpretable modeling, the framework achieves robust, transparent, and adaptable threat detection, paving the way for future advancements in multi-class detection and real-time adaptability. Future efforts will refine the framework by exploring automated tuning methods, such as Bayesian optimization, to enhance performance. Incorporating batch or real-time interpretability techniques will improve scalability and enable self-adaptation to evolving attack patterns through real-time feedback loops. Additionally, the framework will be advanced to detect specific

attack types, including DDoS, infiltration, and port scans, using cutting-edge multi-class threat detection methods. These improvements aim to create AI-driven security systems that are robust, transparent, insightful, and adaptable to the dynamic threat landscape of modern industrial digital infrastructures.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by the Deanship of Postgraduate Studies and Scientific Research at Majmaah University through the project number (R-2025-2136). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Majmaah University: R-2025-2136.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Shailendra Mishra conceived and designed the experiments, performed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Ebtesam Abdulaziz Almutairi conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Reem Alshenaifi analyzed the data, prepared figures and/or tables, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

Raw data is available at GitHub and Zenodo: https://github.com/EbtesamAlmutairi/CNN-LSTM-Intrusion-Detection.git

EbtesamAlmutairi. (2025). EbtesamAlmutairi/CNN-LSTM-Intrusion-Detection: First Release (v1.0). Zenodo. https://doi.org/10.5281/zenodo.17623648

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-cs.3454#supplemental-information.

REFERENCES

Abdelkader S, Amissah J, Kinga S, Mugerwa G, Emmanuel E, Mansour DEA, Bajaj M, Blazek V, Prokop L. 2024. Securing modern power systems: implementing comprehensive strategies to

enhance resilience and reliability against cyber-attacks. *Results in Engineering* **102647**:102647 DOI 10.1016/j.rineng.2024.102647.

Abdullahi SM, Lazarova-Molnar S. 2025. On the adoption and deployment of secure and privacy-preserving IIoT in smart manufacturing. *International Journal of Information Security* **24(1)**:53 DOI 10.1007/s10207-024-00951-8.

Al-Dulaimy A, Jansen M, Johansson B, Trivedi A, Iosup A, Ashjaei M, Galletta A, Kimovski D, Prodan R, Tserpes K, Kousiouris G, Giannakos C, Brandic I, Ali N, Bondi AB, Papadopoulos AV. 2024. The computing continuum: from IoT to the cloud. *Internet of Things* **27(6)**:101272 DOI 10.1016/j.iot.2024.101272.

Al-Quayed F, Ahmad Z, Humayun M. 2024. A situation-based predictive approach for cybersecurity intrusion detection and prevention. *IEEE Access* **12**:34800–34819 DOI 10.1109/ACCESS.2024.3372187.

Aslam MM, Tufail A, Apong RA, De Silva LC, Raza MT. 2024. Scrutinizing security in industrial control systems: an architectural vulnerabilities and communication network perspective. *IEEE Access* **12(1)**:67537–67573 DOI 10.1109/ACCESS.2024.3394848.

Bakhsh SA, Khan MA, Ahmed F, Alshehri MS, Ali H, Ahmad J. 2023. Enhancing IoT network security through deep learning-powered intrusion detection system. *Internet of Things* **24**:100936.

Belarbi O, Khan A, Carnelli P, Spyridopoulos T. 2022. An intrusion detection system based on deep belief networks. In: Wang G, Feng Z, Bhuiyan F, Lu R, eds. *Science of Cyber Security–SciSec 2022 (Lecture Notes in Computer Science).* Vol. 13580. Cham: Springer, 377–396 DOI 10.1007/978-3-031-17551-0_25.

Bobde Y, Narayanan G, Jati M, Raj RSP, Cvitić I, Peraković D. 2024. Enhancing industrial IoT network security through blockchain integration. *Electronics* **13(4)**:687 DOI 10.3390/electronics13040687.

Chen X. 2023. CICIDS2017 and UNBSW-NB15. IEEE Dataport DOI 10.21227/ykpn-jx78.

CIC-IDS. 2017. Dataset. *Available at http://cicresearch.ca/CICDataset/CIC-IDS-2017/Dataset/* (accessed 16 January 2025).

Cisco. 2022. The future of cybersecurity in the IIoT.

El-Sofany H, El-Seoud SA, Karam OH, Bouallegue B. 2024. Using machine learning algorithms to enhance IoT system security. *Scientific Reports* **14(1)**:12077 DOI 10.1038/s41598-024-62861-y.

Holdbrook R, Odeyomi O, Yi S, Roy K. 2024. Network-based intrusion detection for industrial and robotics systems: a comprehensive survey. *Electronics* **13(22)**:4440 DOI 10.3390/electronics13224440.

IBM. 2020. AI for industrial IoT security: protecting your assets.

Iftikhar S. 2024. Cyberterrorism as a global threat: a review on repercussions and countermeasures. *PeerJ Computer Science* **10(6)**:e1772 DOI 10.7717/peerj-cs.1772.

Joha MI, Rahman MM, Nazim MS, Jang YM. 2024. A secure IIoT environment that integrates AI-driven real-time short-term active and reactive load forecasting with anomaly detection: a real-world application. *Sensors* **24(23)**:7440 DOI 10.3390/s24237440.

Khan ZI, Afzal MM, Shamsi KN. 2024. A comprehensive study on CIC-IDS2017 dataset for intrusion detection systems. *International Research Journal on Advanced Engineering Hub (IRJAEH)* **2(2)**:254–260.

Liu J, Hou J, Lin M, Xu Z. 2025. BPSAC: A Blockchain and PUF-Enabled Security Architecture With C-IDS for Maritime Intelligent Transportation Systems. *IEEE Transactions on Intelligent Transportation Systems*.

Mahesh K, Nageswara Rao K. 2025. GRU–CNN hybrid model for intrusion detection in industrial IoT networks. *Indonesian Journal of Electrical Engineering and Computer Science* **39(3)**:1765–1775 DOI 10.11591/ijeecs.v39.i3.pp1765-1775.

Mekala SH, Baig Z, Anwar A, Zeadally S. 2023. Cybersecurity for Industrial IoT (IIoT): threats, countermeasures, challenges and future directions. *Computer Communications* **208(2–3)**:294–320 DOI 10.1016/j.comcom.2023.06.020.

Munirathinam S. 2020. Industry 4.0: Industrial Internet of Things (IIoT). In: Zelkowitz M, ed. *Advances in computers.* Vol. 117. Amsterdam: Elsevier, 129–164.

Nazir A, He J, Zhu N, Qureshi SS, Qureshi SU, Ullah F, Wajahat A, Pathan MS. 2024. A deep learning-based novel hybrid CNN-LSTM architecture for efficient detection of threats in the IoT ecosystem. *Ain Shams Engineering Journal* **15(7)**:102777 DOI 10.1016/j.asej.2024.102777.

NIST. 2021. Cybersecurity framework for IIoT. *Available at https://www.nist.gov/news-events/news/2021/06/nist-announces-new-cybersecurity-framework-industrial-iot*.

Orman A. 2025. Cyberattack detection systems in IIoT networks in big data environments. *Applied Sciences* **15(6)**:3121 DOI 10.3390/app15063121.

Paracha MA, Jamil SU, Shahzad K, Khan MA, Rasheed A. 2024. Leveraging AI for network threat detection—a conceptual overview. *Electronics* **13(23)**:4611 DOI 10.3390/electronics13234611.

Qureshi SS, He J, Qureshi SU, Zhu N, Wajahat A, Nazir A, Ullah F, Wadud A. 2025. Advanced AI-driven intrusion detection for cloud-based IIoT. *Egyptian Informatics Journal* **30**:100644 DOI 10.1016/j.eij.2025.100644.

Sadhwani S, Modi UK, Muthalagu R, Pawar PM. 2024. SmartSentry: cyber threat intelligence in industrial IoT. *IEEE Access* **12**:34720–34740 DOI 10.1109/ACCESS.2024.3371996.

Schummer P, del Rio A, Serrano J, Jimenez D, Sánchez G, Llorente Á. 2024. Machine learning-based network anomaly detection: design, implementation, and evaluation. *AI* **5(4)**:2967–2983 DOI 10.3390/ai5040143.

Shyaa MA, Ibrahim NF, Zainol Z, Abdullah R, Anbar M, Alzubaidi L. 2024. A comprehensive survey on concept drift and feature dynamics in IDS. *Engineering Applications of Artificial Intelligence* **137(20)**:109143 DOI 10.1016/j.engappai.2024.109143.

Srinivasan M, Senthilkumar NC. 2025. Intrusion detection and prevention system (IDPS) for IIoT using a hybrid framework. *IEEE Access* **13(15)**:26608–26621 DOI 10.1109/ACCESS.2025.3538461.

Toussaint M, Krima S, Panetto H. 2024. Industry 4.0 data security: a cybersecurity frameworks review. *Journal of Industrial Information Integration* **39(3)**:100604 DOI 10.1016/j.jii.2024.100604.

Ullah I, Adhikari D, Su X, Palmieri F, Wu C, Choi C. 2024. Integration of data science with the intelligent IoT (IIoT): current challenges and future perspectives. *Digital Communications and Networks* **11(2)**:280–298 DOI 10.1016/j.dcan.2024.02.007.

Wang S, Qureshi MA, Miralles-Pechuán L, Huynh-The T, Gadekallu TR, Liyanage M. 2024. Explainable AI for 6G use cases: technical aspects and research challenges. *IEEE Open Journal of the Communications Society* **5**:2490–2540 DOI 10.1109/OJCOMS.2024.3386872.

Yu J, Shvetsov AV, Alsamhi SH. 2024. Leveraging machine learning for cybersecurity resilience in Industry 4.0: challenges and future directions. *IEEE Access* **12(1)**:159579–159596 DOI 10.1109/ACCESS.2024.3482987.

Zhang H, Huang L, Wu CQ, Li Z. 2020. An effective convolutional neural network based on SMOTE and Gaussian mixture model for intrusion detection in imbalanced dataset. *Computer Networks* **177**:107315 DOI 10.1016/j.comnet.2020.107315.

Zhang J, Zhang X, Liu Z, Fu F, Jiao Y, Xu F. 2023. A network intrusion detection model based on BiLSTM with multi-head attention mechanism. *Electronics.* **12(19)**:4170.

Zhukabayeva T, Zholshiyeva L, Karabayev N, Khan S, Alnazzawi N. 2025. Cybersecurity solutions for industrial internet of things-edge computing integration: challenges, threats, and future directions. *Sensors* **25(1)**:213 DOI 10.3390/s25010213.