AIMS Mathematics, 11(5): 14211–14238.
DOI: 10.3934/math.2026583
Received: 14 February 2026
Revised: 28 March 2026
Accepted: 08 April 2026
Published: 19 May 2026

https://www.aimspress.com/journal/Math

Research article

Mathematical modeling of explainable AI based false data injection attack detection for resilient artificial intelligence of things

Mohammed A. AlAqil[1], Hend Khalid Alkahtani[2], Nasser Allheeib[3], Jahangir Khan[4], Hanadi Alkhudhayr[5], Sami M. Alenezi[6], Malak Bakheet Alharbi[7], Sultan Almutairi[8,*]

[1] Department of Electrical Engineering, College of Engineering, King Faisal University, Al Ahsa, Saudi Arabia
[2] Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia
[3] Department of Information Systems, College of Computer and Information Sciences, King Saud University, Riyadh 11495, Saudi Arabia
[4] Department of Computer Science, Applied College at Mahayil, King Khalid University, Saudi Arabia
[5] Department of Information Systems, Faculty of Computing and Information Technology, King Abdulaziz University, Rabigh 25732, Saudi Arabia
[6] Department of Computer Science, College of Science, Northern Border University, Arar, 91431, Saudi Arabia
[7] Department of Software Engineering, College of Computer Science and Engineering, University of Jeddah, Saudi Arabia
[8] Department of Computer Science, Applied College, Shaqra University, Shaqra 15526, Saudi Arabia

* **Correspondence:** Email: sultan@su.edu.sa.

Abstract: The artificial intelligence of things (AIoT) comprises the integration of artificial intelligence (AI) technologies and the internet of things (IoT) infrastructure. The main aim of the AIoT is to design highly effective IoT operation, enhance human-machine interaction, and improve data analytics and data management. In the AIoT system, the AI is embedded into the infrastructure elements, namely programs and chipsets, which are interlinked via the IoT network. Cybersecurity in the AIoT employs AI technologies, particularly machine learning (ML), deep learning (DL), and neural networks to defend connected IoT devices, networks, and data from recent cyber threats. Since IoT devices have restricted capability, storage, and power, the AI offers intelligent and effective defenses

that can learn to identify anomalies in real time. Therefore, we present a mathematically guided and explainable AI-based framework titled the mutual information-based feature ranking for detecting false data injection attacks (MIFR-DFDIA) approach in resilient distributed cybersecurity networks. The MIFR-DFDIA aims to enhance cybersecurity resilience by accurately identifying and mitigating false data injection attacks that compromise data integrity. Initially, data preprocessing was performed to handle outliers, missing values, and feature standardization for ensuring high-quality input data for analysis. Mutual information was then utilized for optimal feature selection to identify the most informative attributes effectively. For classification, a hybrid model such as a stacked variational autoencoder and a wasserstein generative adversarial network was deployed for robust and precise recognition of false data injection attacks in cybersecurity distributed systems. Finally, the explainable artificial intelligence (XAI) method based SHAP was incorporated to interpret model predictions and improve transparency in decision-making. The experimental result analysis of the MIFR-DFDIA method was carried out for a benchmark dataset, and the comparative analysis exhibited the improved solution over other techniques concerning various metrics.

Keywords: artificial intelligence of things; false data injection attacks; cybersecurity; explainable artificial intelligence; distributed security
Mathematics Subject Classification: 68T07, 68T45

1. Introduction

The integration of information and communication technologies (ICT) into power systems is progressively changing to smart grids [1]. Therefore, applications of power systems, such as state estimation, face major tasks because they rely on telecommunications, an important issue of their susceptibility to cyber-threats [2]. Opponents of power grids could connect and operate system variable measurements, tackling the measurement gadgets and cooperating with the communication systems [3]. Accordingly, compromised system conditions can affect the grid operation, resulting in economic and manual effects on the power system. Among ordinary attacks in cyber-manual methods, the false data injection attack (FDIA) is a complex threat to the state estimation. Unlike other attacks, namely jamming and distributed denial of service (DDoS), an effective FDIA can avoid the traditional residual-driven bad information recognition module [4]. Without an advanced detection module, FDIA can be secretly launched several times, rendering an important attack on the grid. The significance of these threats is profound, impacting not only the values but also consumers who pay for continuous electrical power [5]. Additionally, it is extremely important to counter FDIA before the evil invasion grows craftier and more decisive.

With advancements in artificial intelligence (AI) techniques, namely machine learning (ML) and deep learning (DL), there is an essential focus on detecting FDIA [6]. Moreover, information-driven detection approaches are accepted as an effective tool for attack detection on smart grids and power systems [7]. These solutions are normally based on statistical models or ML to conclude a method of the system from measurement signals and past information. In addition, with the accessibility of progressive computation resources and their effective application in domains, several scholars have examined the likelihood of employing DL techniques for the detection of FDIAs [8]. Furthermore, DL sends an enhanced method to classical (ML) solutions. When there is adequate information, DL models

nearly send exceptional outcomes. Therefore, DL techniques have been gradually applied to address the cybersecurity problem in relation to other domains like software vulnerability, image processing, and NLP [9]. Even though the AI method, particularly DL and ML, can offer imposing performances on benchmark databases in many cyber security filed applications, namely spam e-mail filtering, malicious application identification, Intrusion detection, fraud detection, Botnet detection, and FDIA detection, they may perform errors due to their black box nature, and a few of which are costly compared to traditional cyber defensive methods [10].

There is a rise of explainable artificial intelligence (XAI) as a transformative method in cybersecurity [11]. XAI is a collection of methods and processes that enable individual users to interpret, understand, and trust the outcomes made by ML methodologies. By shedding light on the "black-box" behavior of AI methods, XAI improves transparency and offers a window into the reasoning behind an AI-based decision. In the context of FDI attack detection, XAI enables security forecasters to obtain crucial insights into the factors and information features that run the method for identifying a particular event as malicious [12]. Besides this, it builds confidence in the automatic detection method but also enables a more contextual and informed comprehension of attacks, assisting human experts to authenticate findings and correct the effective method [13]. Figure 1 represents the general process of FDIA.

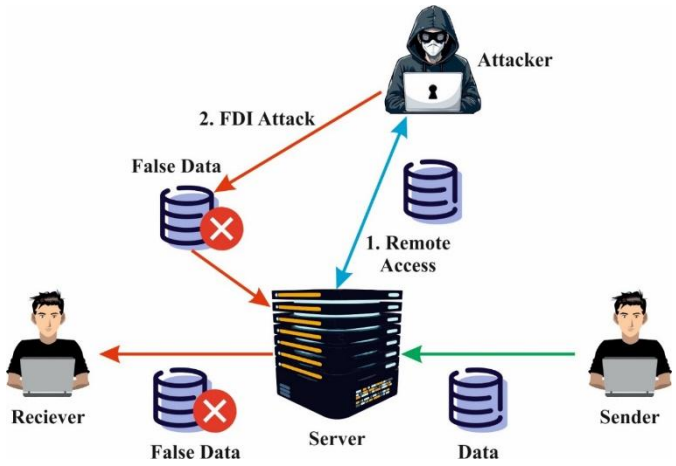

Figure 1. Process of FDIA.

1.1. Key contributions and novelty

To address the increasing security challenges in distributed cybersecurity networks caused by FDIA, we present a mathematically guided and explainable AI-based framework titled Mutual Information-Based Feature Ranking for Detecting False Data Injection Attacks (MIFR-DFDIA). In the proposed model, we aim to effectively detect and mitigate FDIA by employing robust feature selection, hybrid DL, and explainable methods. We follow a structured process, including data preprocessing, feature selection, classification, and interpretation. The efficiency of the proposed MIFR-DFDIA approach is assessed using benchmark datasets under different evaluation metrics. The major contributions of this manuscript are as follows:

- Utilize a mutual information (MI) based feature selection approach to identify the most relevant features, helping to enhance detection accuracy while minimizing unnecessary data.

- Design a hybrid deep learning model combining stacked variational autoencoder (SVAE) and wasserstein generative adversarial network (WGAN) for effective and reliable detection of FDIA in distributed cybersecurity environments.
- Incorporate a SHAP-based explainable AI technique to interpret model predictions and emphasize critical features contributing to attack detection decisions, thereby improving transparency in decision-making.
- Assess the proposed model on benchmark datasets and demonstrate the improved performance of the MIFR-DFDIA model compared to models across metrics.

Novelty: The novelty of this work lies in combining feature selection, hybrid deep learning, and explainability in a single framework for FDIA detection. Unlike existing techniques, it focuses on accuracy and interpretability together. Additionally, the model is designed to handle real-world cybersecurity scenarios more effectively.

1.2. Article layout

This manuscript proceeds as follows: In Section 2, we present a review of the related studies in the context of FDIA detection. In Section 3, we detail the methodology leveraged in this study, comprising the design of the detection technique, the procedure of feature selection, the classification process, and explainability analysis. In Section 4, the experimental validations are given, comprising a detailed examination of performance through several metrics. Last, in Section 5, conclusions are outlined.

2. Review of related works

In this section, a literature analysis on FDIA detection in distributed cybersecurity networks is presented. Soni et al. [14] introduced an efficient method for FDI attack identification and detection. Feature selection employing SHAP was presented for increasing the effectiveness of ML methods. To improve the interpretability of the structure, a combined local interpretable model-agnostic explanations (LIME)-driven XAI model was constructed, offering transparency in attack detection. There is a requirement for an effective method that may classify and detect this attack, ensuring the proper methods of autonomous vehicles (AV). Jain et al. [15] suggested a hybrid ML structure that integrates deep feature extraction employing CNNs with the strong classification abilities of XGBoost. Moreover, to ensure interpretability that is crucial in security-sensitive application, combine an XAI method utilizing SHAP values. These explanations assist in classifying the major characteristics that influence every classification decision, raise trust, and support informed activities. Ji et al. [16] proposed an FDIA detection system that depends on DL with multiscale feature fusion. Primarily, the improved CNN (ICNN) is employed for predicting measurement information by integrating CNN with the Inception v1 mechanisms. Therefore, the attention module is proposed in the ICNN to extract and fuse full and partial characteristics of measurement information. By appropriately applying the function among state vectors and measurements, the state information is made with forecast measurement information.

Feng et al. [17] introduced an innovative scheme for adversaries known as the SMAN. For higher adaptation to harsh learning states with only some labeled examples, the system integrates a semi-

supervised module with an advanced GAN. The discriminator and generator are intended to optimize the method for better detection precision in semi-supervised training. Alamro et al. [18] proposed a modified red fox optimization with DL enabled FDIA detection (MRFODL-FDIAD) in the CPPS setting. This method mostly classifies and detects FDIAs in the CPPS setting. It includes a 3-phase procedure, such as hyperparameter tuning, detection, and preprocessing. For FDIA detection, the presented model employs a multi-head attention-driven LSTM (MBALSTM) method. Huang et al. [19] introduced a DRL in the FDIAs detection approach. They concentrated on state attention for solving the issues to extract state features in the present RL detection approach. Moreover, this method adds an attention module to the method-free DRL detection method of the state feature extraction. Selective attention assists in concentrating on the significant role instead of being distracted by unrelated details, creating states that are more distinguishable and representative, thus enhancing the RL structure for detecting attacks quickly and precisely.

The researchers in [20] presented a remedial action system (RAS) that depends on the idea of DL for lessening the effects of FDIA cyber-attacks on an intelligent power system. In the intention of the RAS, LSTM are combined into a D-RNN for efficiently processing the information of an IAF, classifying the desired reaction module. Guo et al. [21] examined the security problem of the multi-sensor remote estimation method, an optimum stealth FDIA system that depends on past and present residuals that only deal with the measurement residuals of partial sensors.

Zhou et al. [22] developed an attention-based interaction-aware spatio-temporal graph neural network (GNN) for forecasting pedestrian trajectories. Dual elements are presented in this method such as the spatial graph neural network for interaction modeling and temporal GNN for motion feature extraction. Fan et al. [23] introduced an innovative FS architecture for semisupervised and unsupervised approaches. To create effective usage of data distribution to assess features, the architecture integrates the data structure learning and FS in a combined design. Yang et al. [24] developed a phased AC FDIA targeting in the generation rescheduling and load shedding. After providing the false data into the evaluation, the assessed conditions can be departed from those in standard states. This method is used to extract the spatial and spectral features of the modes decomposed from the projected states by employing variational mode decomposition (VMD).

Tian et al. [25] implemented a great perception for the security risks of DL-based multi-label FDIA detectors, the two alternating direction method of multipliers (ADMM) based adversative attacks that can be appropriate to dual diverse conditions, thus generating the attacks more accurate and viable. Tian et al. [26] inspected the multi-label adversarial sample attacks corresponding to multi-label FDIA locational detectors and used the wide-ranging multi-label adversarial attack method, such as the muLti-labEl adverSarial falSe data injectiON (LESSON) attack. The LESSON attack technique comprises 3 major architectures that must be supported to determine appropriate multi-label adversarial perturbations in the physical restrictions to avoid the neural attack location (NAL) and bad data detection (BDD). Table 1 demonstrates a review on FDIA attacks.

Table 1. Overview of detecting false data injection attacks.

References	Techniques	Datasets	Outcomes	Goals
Soni et al. [14]	SHapley additive exPlanation (SHAP) and Machine learning models	FDI attack dataset	Accuracy of 99%	To present an effective model for detecting and classifying FDI threats and to contribute to the reliability and security of individual operations.
Jain et al. [15]	Convolutional neural network (CNN) and explainable AI (XAI) models	CIC IoT2024-DIAD dataset	Accuracy of 99.92%	To project a hybrid model for IoT gadget recognition by integrating NN technique and categorization methodologies.
Ji et al. [16]	Improved CNN model	NYISO dataset	Accuracy of 98.38%	To suggest an FDIA recognition approach that depends on learning models with multi-scale feature fusion.
Feng et al. [17]	Adversarial network	Huge dataset	Accuracy of 92.26%	To examine a graph attention-based model for enhancing generators and increasing the dependability of chosen instances.
Alamro et al. [18]	Multi-head attention-based LSTM (MBALSTM) and modified red fox optimizer (MRFO)	CPPS dataset	Accuracy of 96.78% and 96.68%	To investigate an innovative approach for classifying and detecting the FDIA elements required to achieve to dependability, which are extremely prone to threats.
Huang et al. [19]	Deep Reinforcement Learning model	NA	NA	To investigate a method that adds an attention module to model-free recognition approach for eliminating state attributes.

Continued on next page

Naderi and Asrari [20]	Deep learning approaches	Attack dataset	Threat rate reduced by 30%	To recommend a RAS and lessen the impacts of FDI cyber threats on the power system.
Guo et al. [21]	Residual based network models.	Attack dataset	-	To intend the stealthy threat approach extremely degrades the estimated outcome of a multi-sensor system with confined threat resources.
Zhou et al. [22]	Spatial GNN and CNN models	ETH and UCY datsets	Effective threat mitigation	To integrate spatio-temporal features for trajectory forecasting.
Fan et al. [23]	Semi-supervised FS method	Attack dataset	Achieves superior outcomes	Select features that capture global and local data distributions.
Yang et al. [24]	LSTM-autoencoder approach	IEEE 14 and 118-bus systems datasets	Enhanced system security	To identify attacks targeting generation rescheduling and load shedding.
Tian et al. [25]	Deep learning algorithms	Attack dataset	Better attack detection	To evaluate adversarial attack strategies under different operational scenarios.
Tian et al. [26]	CNN and DNN methods	IEEE 14-bus, 30-bus, and 118-bus systems	Reduced false alarms	To highlight critical security risks in deep learning-based FDIA locational detection.

Although researchers focus mainly on enhancing FDIA detection accuracy, they do not handle dynamic and real-time network conditions effectively. Most of the models rely on complex deep learning and hybrid models, which increase computational cost and reduce practical usability. Although some methods include explainability techniques like SHAP and LIME, they do not fully address model efficiency and scalability. Several models are designed for specific applications, like power systems, constraining their generalization to other distributed network environments. Moreover, feature selection is not properly optimized in many studies, resulting in redundant and less informative features. Hence, there is a need for a more efficient and generalized framework for reliable FDIA detection in distributed cybersecurity networks. Therefore, the proposed model focuses on the design of a mathematically guided and explainable AI-based approach for identifying and classifying FDIA attacks in distributed cybersecurity networks, aiming to ensure reliable and secure data transmission.

3. Methodological framework

The proposed MIFR-DFDIA model follows a systematic multi-stage pipeline designed for accurate and transparent detection of FDIA in resilient distributed cybersecurity networks. Figure 2 demonstrates the conceptual workflow of the MIFR-DFDIA system. The proposed model involves several steps, as illustrated in Figure 2.

Step 1: Data Preprocessing: Raw data are collected and preprocessed through outlier handling, missing value imputation, and feature standardization to guarantee high-quality, consistent input data for precise attack detection.

Step 2: Mutual information-based feature selection: The preprocessed data are then passed to the MI method, which identifies and selects the most informative and non-redundant attributes by measuring the dependency among each feature and the target class, thereby minimizing dimensionality and improving detection robustness.

Step 3: Hybrid deep learning-based classification: The selected features are fed into a SVAE and WGAN model, which learns complex data distributions and latent representations for accurate classification of FDIAs in distributed cybersecurity environments.

Step 4: Explainable AI Interpretation: Finally, the SHAP-based XAI technique is integrated to provide instance-level and global interpretability, which emphasizes the most significant features contributing to each prediction and enhances the transparency and trustworthiness of the attack detection decisions.

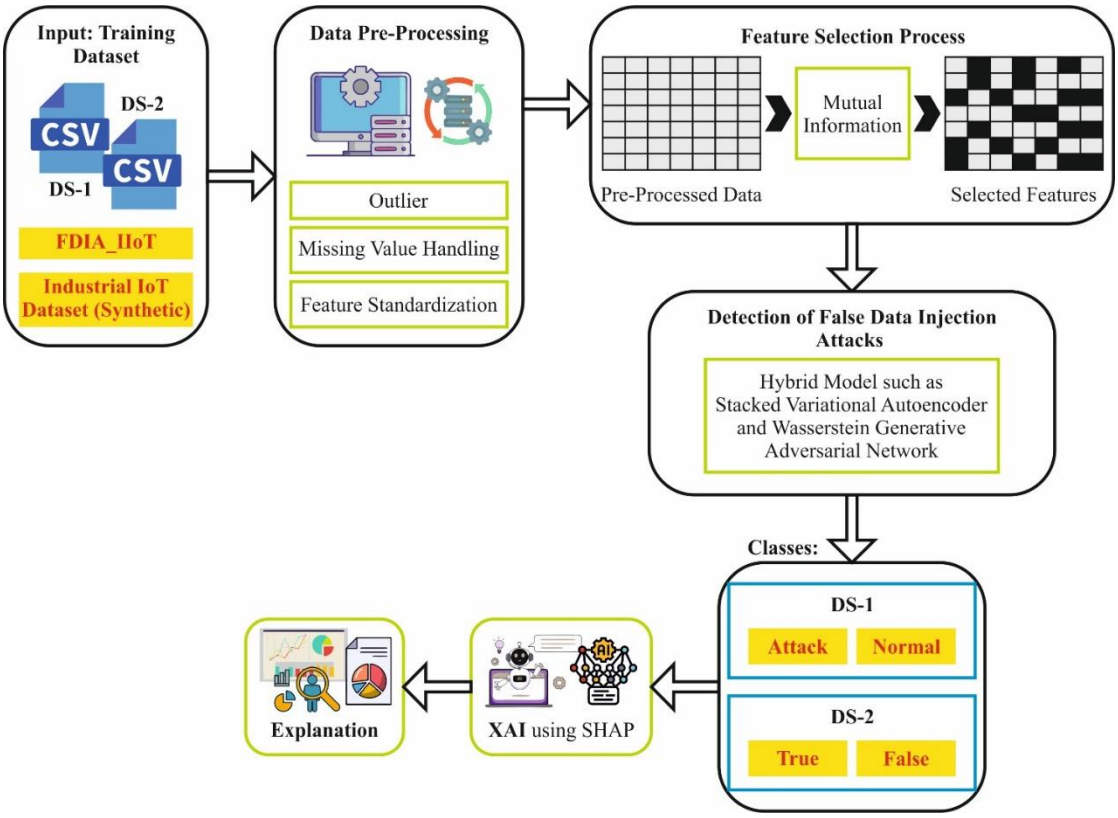

Figure 2. Overall architecture of the MIFR-DFDIA model.

3.1. Data preprocessing

Initially, data preprocessing is used to handle outliers, missing values, and feature standardization for high-quality input data.

Outlier and missing value handling: Data preprocessing is executed to employ a systematic model. Primarily, outliers are recognized and eliminated to employ the method of SD. Then, missing values are managed to employ mean interpolation.

$$x_i = \frac{1}{4}(x_{i-2} + x_{i-1} + x_{i+1} + x_{i+2}). \tag{3.1}$$

Feature standardization: Each input variable is standardized to remove training bias of the model caused by dimensional modifications in multi-source attributes:

$$x_{std} = \frac{x-\mu}{\sigma}. \tag{3.2}$$

Here, μ and σ refer to the mean and SD of the training set attributes, and x signifies the original feature value, correspondingly. This transformation modifies the feature distribution to zero mean and unit variance, substantially enhancing the adaptability of the model to heterogeneous aspects. The validation and test sets are strictly standardized based on the statistics (mean and SD) of the training set to lessen the risk of data leakage. This normalizes the scale of every feature variable and lays a reliable data foundation for the succeeding model.

3.2. Feature reduction using mutual information

In this section, mutual information (MI) is leveraged as a dimensionality reduction model to effectually obtain the most relevant features. Initially, MI is employed for ranking the features [27]. ML methods are employed for evaluating these combinations and classifying the most significant feature set depending on the accuracy metric. The roles of MI and ML approaches here are to employ MI for a clear, interpretable basis for ranking features, with ML paradigms defining the final subset from experimental outcomes. This method supports fairness and employs a personalized optimizer during selection. MI develops as an essential metric, measuring the relevance among features and target variables. It computes a level to which the information of a feature variable reduces vagueness regarding the target variable. A more extensive MI score specifies a more distinct relationship among the target and a feature outcome. The formulation for computing the MI among the target variable V and feature variable U is expressed by

$$MI(U;V) = \sum_{v\in V}\sum_{u\in U} p(u,v)\log\left(\frac{p(u,v)}{p(u)p(v)}\right), \tag{3.3}$$

whereas $p(u,v)$ denotes a joint probability distribution of u and v, and $p(u)$ and $p(v)$ imply a marginal probability distribution of u and v, respectively. Features revealing higher MI with target variables is considered effective because they specify robust numerical dependency. An MI-driven FS approach is used, where features are ranked depending on their MI scores regarding FDIA labels. An adaptive threshold (e.g., top-k features or $\geqslant$90th percentile MI score) is applied to retain the most beneficial features, guaranteeing relevance to FDIA patterns while minimizing redundancy and enhancing model generality.

3.3. FDIA detection algorithm

In this step, SVAE and WGAN are utilized for precise detection of FDIA in cybersecurity distributed systems. The SVAE-WGAN framework is particularly intended for capturing the fundamental structure of intricate data distribution and making higher-quality information with improved reliability and range [28].

The SVAE applies a dual-layer stacked VAE structure, which is intended for balancing the method's expressive power and training intricacy efficiently. In an initial layer, the model takes an early latent feature of input data, whereas the secondary layer takes out hidden abstract representation, thus improving the method's capability to model an intricate data structure. This hierarchic model alleviates the restrictions of an inadequate feature model in one layer VAE and decreases an overfitting hazard and training uncertainty. The multi-layer framework enhances the method's ability for managing a higher-dimensional and intricate data, enabling the group of representative data models. When equated to conventional GAN, the training procedure is more even, and the generated models display better exposure. Conversely, SVAE trusts on error of reconstruction and KL divergence as an optimization criterion that appear in the absence of feature and accuracy when producing higher-dimensional information. The Wasserstein distance delivers an even and significant measure of dissimilarity among real and generated samples. However, WGAN training needs common upgrades to discriminators, enlarging computation overhead. Furthermore, the training of the generator has been deeply reliant on the discriminator, which might influence convergence constancy. To incorporate the powers of WGAN and SVAE, the SVAE-WGAN method efficiently tackles the restrictions of SVAE in making higher-quality models. The module of SVAE ensures consistency and range averting mode collapse, whereas the WGAN module improves model quality over accurate response and a more effectual training procedure.

The SVAE-WGAN model deploys SVAE capability for capturing hidden data feature and WGAN's generation abilities. This united model presents major benefits in tackling the tasks of intricate data modeling.

In the SVAE method, an input layer handles raw information over denoising or normalization models, ensuring an even data distribution and alleviating an impact of noise. The hidden layers (HLs) include dual stacked VAEs that gradually map the information into hidden spaces, removing hierarchic attributes to improve the ability for processing intricate data structures. It averts problems like inadequate feature representation. The decoding rebuilds the data layer-wise, incorporating the hidden space representation to enhance the superiority and range of generated data. The layer of output enhances the encoding and decoding by enlarging the probability of reconstructed data and reducing the divergence of Kullback-Leibler (KL), thus improving the complete efficacy and representation ability of the method. This ensures strong performance in seizing complex data patterns while upholding constancy throughout training. Figure 3 shows the architecture of SVAE.

In the primary layer of VAE encoding, the encoding obtains an inputting data x, and mapping it to the latent variables $y_1.y_1$ causes the hidden variable to obey a Gaussian distribution with variance $\sigma_1^2(x)$ and mean $\mu_1(x)$.

$$q_{\varphi_1}(y_1|x) \sim \mathcal{N}\big(y_1; \mu_1(x), \sigma_1^2(x)\big). \tag{3.4}$$

The decoding creates models from the hidden region y_1 and makes the recreated inputting

data $\hat{x}_1$, which obeys a Gaussian distribution

$$p_{\theta_1}(x|y_1) \sim N(x; \hat{x}, \sigma^2). \tag{3.5}$$

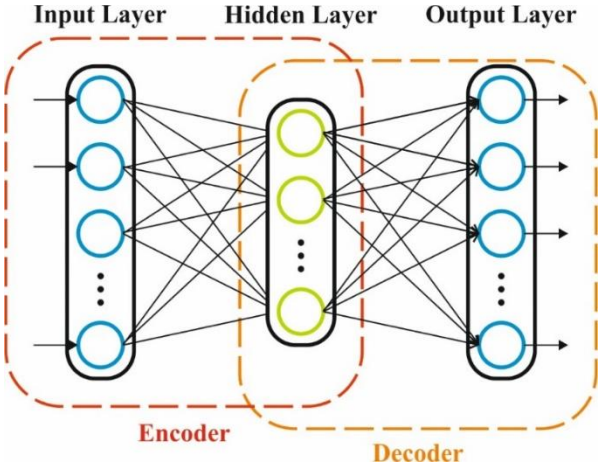

Figure 3. Architecture of SVAE.

In the secondary layer encoding, the initial layer latent variables y_1 are employed as an input to produce the secondary layer latent variable y_2. They also obey a Gaussian distribution with variance $\sigma_2^2(y_1)$ and mean $\mu_2(y_1)$.

$$q_{\phi_2}(y_2|y_1) \sim N\big(y_2; \mu_2(y_1), \sigma_2^2(y_1)\big). \tag{3.6}$$

From y_2, the second layer produces the reconstructed latent variable $\hat{y}$ and then passes over the first layer decoding for rebuilding an original input data $\hat{x}$.

$$p_{\theta_2}(y_1|y_2) \sim N(y_1; \hat{y}_1, \sigma^2). \tag{3.7}$$

The complete loss function is stacked as VAE and composed into dual portions:

(1) The reconstruction loss signifies the errors of both layers.

$$L_{recon} = -\mathbb{E}_{q_{\phi_1}(y_1|x)}\big[\log p_{\theta_1}(x|y_1)\big] - E_{q_{\phi_2}(y_2|y_1)}\big[\log p_{\theta_2}(y_1|y_2)\big]. \tag{3.8}$$

(2) The loss of KL divergence enumerates the dissimilarity among the hidden distributions made by the encoding and the previous distribution.

$$D_{KL}(q_{\phi_1}(y_1|x)||p(y_1)) + D_{KL}(q_{\phi_2}(y_2|y_1)\big|\big|p(y_2)\big). \tag{3.9}$$

Therefore, the complete loss function of SVAE is signified below:

$$L_{SVAE} = L_{recon} + D_{KL}. \tag{3.10}$$

In WGAN, the generator's objective is to enlarge the score of discriminators, causing the generated sample to be close to the distribution of true samples. The function of loss comprises an

output difference among samples of those generated and real.

$$L_{SVAE\text{-}WGAN} = L_{SVAE} + \lambda\left(\mathbb{E}_{x_r\sim p_{data}}[D(x_r)] - E_{y\sim p_y}\left[D\left(G(y)\right)\right]\right). \tag{3.11}$$

whereas y means a noise input, x_r signifies the real data, and λ denotes a parameter, which balances SVAE and WGAN losses. $G(y)$ represents data generated by a generator

$$L_G = -E_{y\sim p_y}\left[D\left(G(y)\right)\right]. \tag{3.12}$$

SVAE-WGAN unites the SVAE variational implication with the WGAN's generative adversarial training. The SVAE enhances the encoding and decoding by minimalizing the KL divergence and error of reconstruction, whereas the WGAN enhances the generator by minimalizing the loss of a discriminator and generator, causing generated data to be closer to the distribution of real data. The classification network can be implemented using

$$\hat{z} = Softmax(W \cdot y + b). \tag{3.13}$$

3.4. XAI-based analysis

Last, for the XAI technique, SHAP is employed to interpret model predictions and enhance transparency in decision-making. The feature analysis in this work is examined utilizing the SHAP based study [29]. The SHAP-based analysis helps in calculating the contribution of all features to prediction. The usage of SHAP in feature engineering is to predict the gap between interpretability and model performance. The resultant equation for the SHAP-based feature study is provided in Eq (3.14):

$$\phi_x = \sum_{p\subseteq F\{x\}} \frac{|p|!(k-|p|-1)!}{k!}\left[f\left(p\cup\{x\} - f(p)\right)\right]. \tag{3.14}$$

Here, p defines the feature's subset exclude $x(p \subseteq F\{x\})$, and $|p|$ describes the feature counts within the subset p. $f(p)$ epitomizes the features in x that gives prediction. $f(p \cup x)$ represents the prediction over features in p and x. The model output corresponding to the input is stated as the number of the Shapley values, and the same is presented in Eq (3.15):

$$f(x) = f(\phi) + \sum_{x=1}^{k} \phi_x. \tag{3.15}$$

The feature significance in creating the last assessment is evaluated by calculating the mean absolute Shapley value for all features through each sample, as shown below:

$$s_f = \frac{1}{n}\sum_{i=1}^{n} |\,\phi_x^i|. \tag{3.16}$$

Here, φ_x^i signifies the Shapley value of feature x for the i^{th} sample. n describes the sample counts. The features that are unique in the decision process of fault detection are offered. The Shapley dependency graphs are excellent at offering a visual depiction of feature dependency and their contribution for modeling predictions. The negative and the positive dependency among the features is represented utilizing dependency graphs. Figure 4 depicts the general diagram of the XAI algorithm.

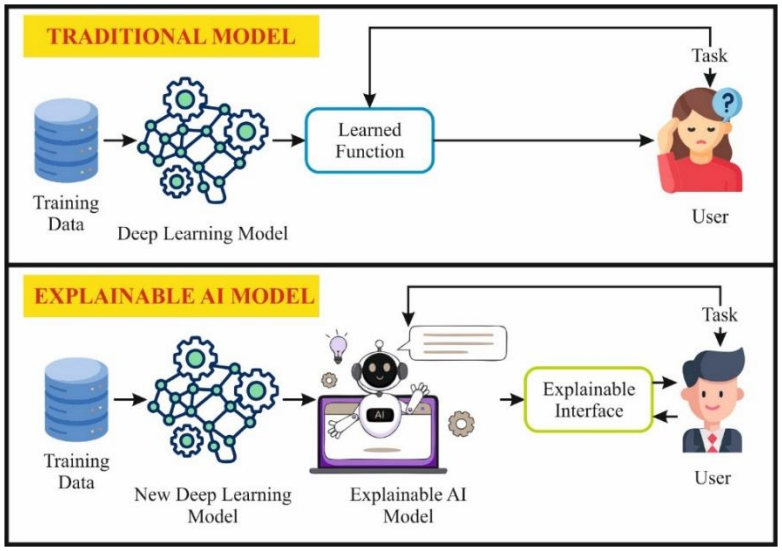

Figure 4. Architecture of XAI.

3.5. Performance evaluation metrics

To calculate the performance of the model, 4 standard measures are utilized, given below. To describe a positive instance labeled with an enhanced learning outcome, and a negative instance with a lesser outcome depends on error metrics and system feedback scores. Accuracy equates the complete accuracy of a method as a proportion of true predictions to the total number of predictions. Precision evaluates the percentage of true positives predicted amid positives, whereas recall calculates the fraction of actual positives appropriately categorized. The F1-score measures the harmonic mean of precision and recall, balancing both measures. These metrics offer a wide-ranging assessment of outcomes.

$$Accuracy = \frac{TP+TN}{Total\ Samples}, \tag{3.17}$$

$$Precision = \frac{TP}{TP+FP}, \tag{3.18}$$

$$Recall = \frac{TP}{TP+FN}, \tag{3.19}$$

$$F_1 = \frac{2\times Precision\times Recall}{Precision+Recall}. \tag{3.20}$$

Here, TP implies the number of positive samples appropriately categorized as positive; TN implies the number of negative samples properly categorized as negative; FP depicts the number of negative examples inappropriately identified as positive; and FN implies the number of positive samples improperly identified as negative.

4. Results and discussions

In this paper, different experiments are carried out for varying training configurations, namely 500, 1000, 1500, 2000, 2500, and 3000 epochs. The proposed approach is simulated using Python 3.8.5 tool on PC i5-8600k, GeForce 1050Ti 4GB, 16GB RAM, 250GB SSD, and 1TB HDD. The

parameter settings are provided as follows: learning rate: 0.01, dropout: 0.5, batch size: 5, and activation: ReLU. For experimental validation, 70% of the training and 30% of the testing datasets are implemented. The performance assessment of the MIFR-DFDIA model is examined under the False Data Injection Attack dataset. It has dual datasets such as the FDIA-IIoT [30] and IIoT [31] datasets. The description and outcomes of these dual datasets are explained below.

4.1. Dataset description

FDIA-IIoT dataset: The FDIA-IIoT dataset includes 14,925 samples, split into 8,688 "Normal" instances and 6,237 "Attack" instances, offering a relatively balanced class distribution for learning discriminative patterns between legitimate and malicious behaviors. This dataset captures different false data injection attack scenarios in IIoT environments, reflecting realistic sensor-level manipulation and integrity violations commonly observed in cyber-physical infrastructures. Out of 30 original features, 21 informative features are selected, which helps reduce computational overhead and enhances detection efficiency under resource-constrained edge computing environments. This dataset serves as a benchmark to evaluate the robustness of the proposed approach.

IIoT Dataset: The IIoT dataset encompasses 20,000 samples with an equivalent distribution of 10,000 "TRUE" and 10,000 "FALSE" samples, guaranteeing a fully balanced binary classification setting for unbiased performance evaluation. The dataset is synthetically generated to simulate factory sensor measurements and operational states in IIoT environments, enabling controlled experimentation of attack detection mechanisms under diverse operating conditions. This demonstrates multivariate sensor correlations and noise patterns that resemble real-world IIoT deployments in smart manufacturing and cyber-physical systems. Although the dataset originally contained 21 features, only 17 discriminative features are selected to improve model compactness and real-time detection performance. This dataset supports the validation of scalability and generalization of the proposed detection model across IIoT data distributions.

4.2. Result analysis of the FDIA-IIoT dataset

Figure 5 demonstrates the correlation matrix created by the MIFR-DFDIA model for the FDIA-IIoT dataset. The outcomes denote that the MIFR-DFDIA approach proficiently classifies each class.

Table 2 and Figure 6 describe an attack detection of the MIFR-DFDIA technique for the FDIA-IIoT dataset under different epochs. For epoch 500, the proposed MIFR-DFDIA model obtains an average $accur_y$ of 97.37%, $preci_n$ of 97.41%, $recal_l$ of 97.37%, and $F1_{score}$ 97.39%, respectively. Additionally, for epoch 1500, the proposed MIFR-DFDIA approach attains an average $accur_y$ of 98.16%, $preci_n$ of 98.02%, $recal_l$ of 98.16%, and $F1_{score}$ of 98.09%. Moreover, for epoch 2500, the proposed MIFR-DFDIA method reaches an average $accur_y$ of 98.27%, $preci_n$ of 98.14%, $recal_l$ of 98.27%, and $F1_{score}$ of 98.21%.

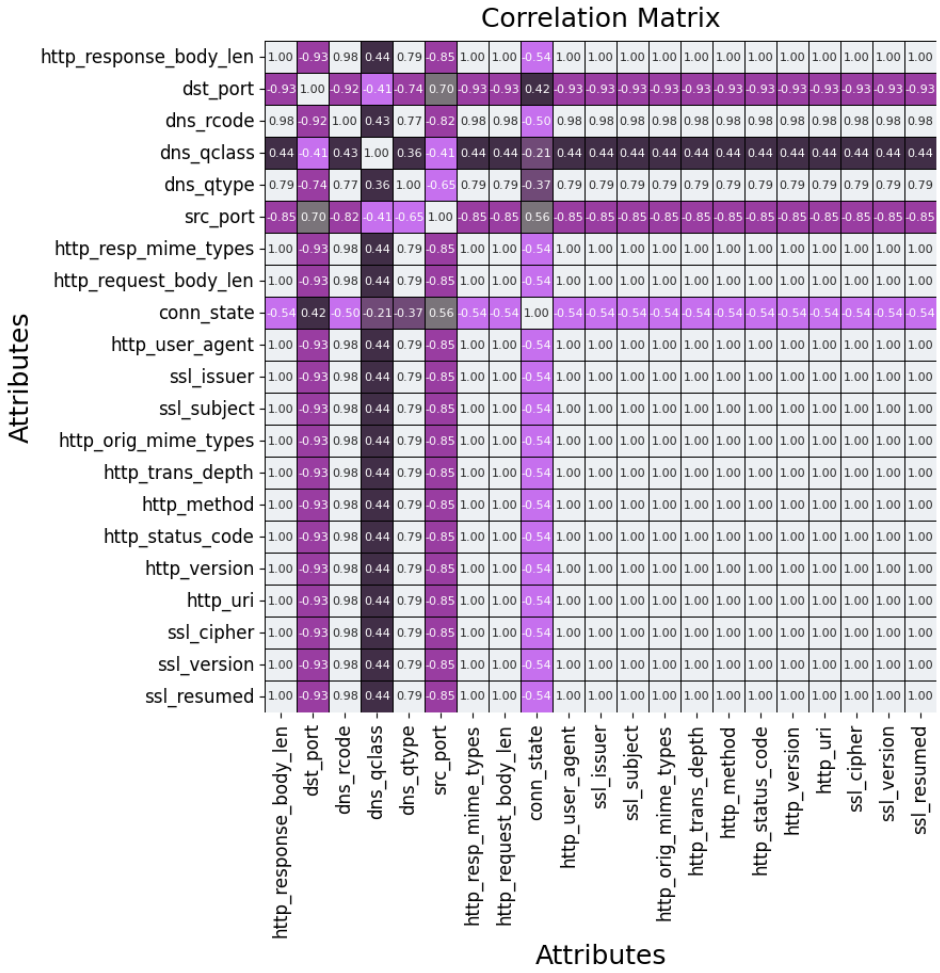

Figure 5. Correlation matrix of the selected features for the FDIA-IIoT dataset.

Table 2. Attack detection of the MIFR-DFDIA model for the FDIA-IIoT dataset.

Training cycle and overall	$Accur_y$		$Preci_n$		$Recal_l$		$F1_{score}$	
	Attack	Normal	Attack	Normal	Attack	Normal	Attack	Normal
500 Learning Epochs	96.83	97.92	97.09	97.73	96.83	97.92	96.96	97.82
Average	97.37		97.41		97.37		97.39	
Epoch-1000	97.95	98.15	97.43	98.52	97.95	98.15	97.69	98.33
Average	98.05		97.98		98.05		98.01	
Epoch-1500	98.3	98.02	97.27	98.77	98.3	98.02	97.78	98.39
Average	98.16		98.02		98.16		98.09	
Epoch-2000	98.08	98.22	97.53	98.61	98.08	98.22	97.8	98.41
Average	98.15		98.07		98.15		98.11	
Epoch-2500	98.4	98.15	97.44	98.84	98.4	98.15	97.92	98.49
Average	98.27		98.14		98.27		98.21	
Epoch-3000	98.91	99.18	98.86	99.22	98.91	99.18	98.89	99.2
Average	99.05		99.04		99.05		99.04	

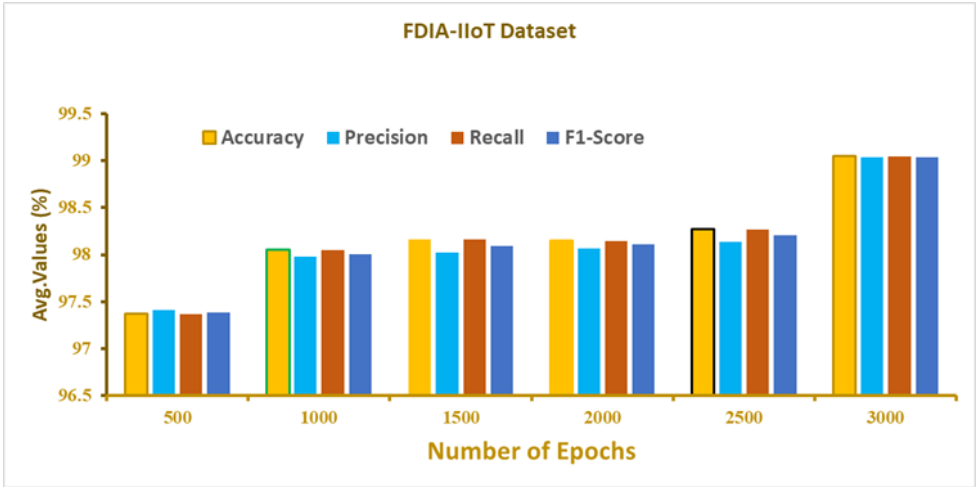

Figure 6. Average values of the MIFR-DFDIA model for the FDIA-IIoT dataset.

Figure 7 displays the training (TRAN) and validation (VALD) accuracies of the MIFR-DFDIA methodology for the FDIA-IIoT dataset over 3000 epochs. Both curves progressively surge and gradually converge, specifying that the technique is proficiently learning. The VALD accuracy reliably remains better than the TRAN accuracy, inferring that the approach is not over-fitting and generalizes well to unnoticed data. The variations in accuracy are because of the task intricacy, but the overall rising tendency indicates robust performance and model stability.

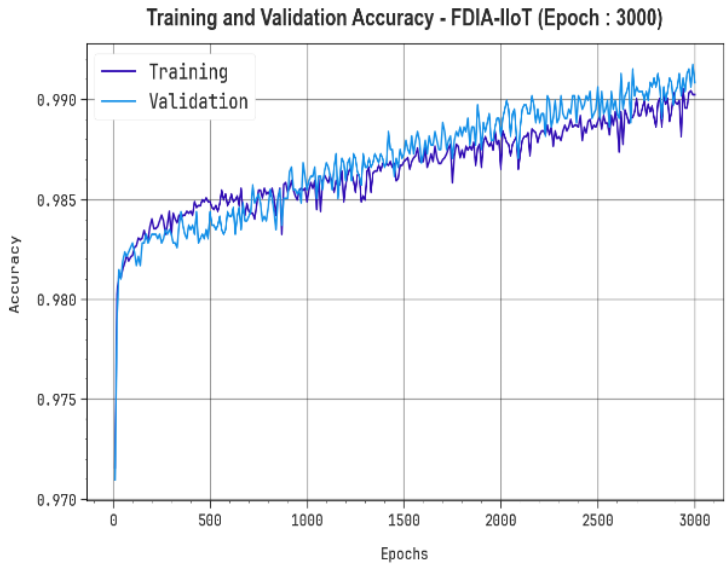

Figure 7. Accuracy curve of the MIFR-DFDIA algorithm for the FDIA-IIoT dataset.

Figure 8 illustrates the loss of MIFR-DFDIA technology for the FDIA-IIoT dataset over 3000 epochs, as measured by TRAN and VALD. Both curves show reliable downward tendencies, proving that the method efficiently reduces error in learning. The VALD loss stays somewhat inferior to the training loss across most epochs. Even though some variations are observed, it has become balanced and more sustained.

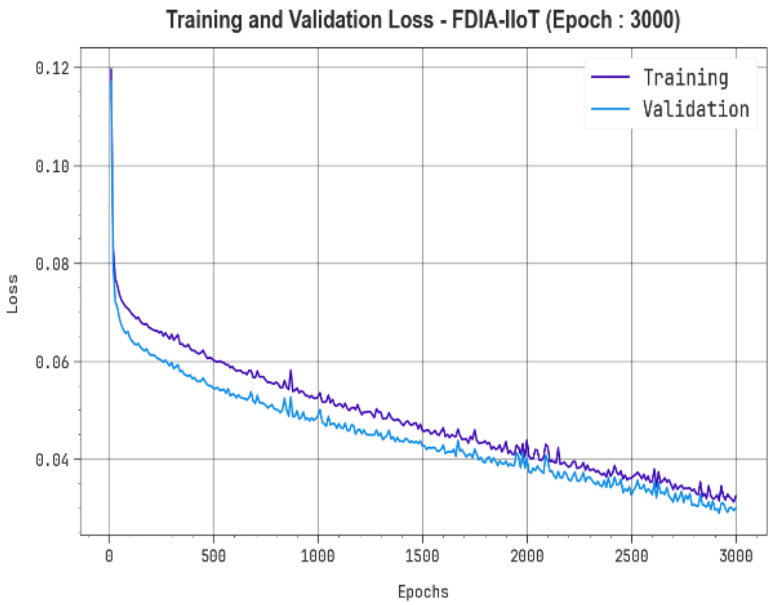

Figure 8. Loss curve of the MIFR-DFDIA algorithm for the FDIA-IIoT dataset.

Table 3 and Figure 9 offer a performance comparison of the MIFR-DFDIA technique for the FDIA-IIoT dataset with the system under various metrics [32–35]. The presented MIFR-DFDIA method attains the highest $accur_y$ of 0.9904, whereas the LR models, KNN algorithm, XGBoost, ANN algorithm, FedIFD, FL-BERT, MIC-XGB, VMD-ELM, MTMCN, and extra trees technologies have the lowest $accur_y$ of 0.6202, 0.8457, 0.9764, 0.8000, 0.9700, 0.8400, 0.9798, 0.9000, 0.9830, and 0.9400, respectively. Likewise, based on $F1_{score}$, the presented MIFR-DFDIA method achieves a maximal $F1_{score}$ of 0.9904%, whereas the LR models, KNN algorithm, XGBoost, ANN algorithm, FedIFD, FL-BERT, MIC-XGB, VMD-ELM, MTMCN, and extra trees models attain a minimal $F1_{score}$ of 0.9424, 0.9498, 0.7320, 0.9222, 0.8362, 0.7597, 0.9763, 0.9043, 0.8968, and 0.9122, respectively.

Table 3. Comparative analysis of the MIFR-DFDIA system for the FDIA-IIoT dataset.

Techniques	$\boldsymbol{Accur_y}$ (*100%)	$\boldsymbol{Preci_n}$ (*100%)	$\boldsymbol{Recal_l}$ (*100%)	$\boldsymbol{F1_{score}}$ (*100%)
LR models	0.6202	0.6926	0.9341	0.9424
KNN algorithm	0.8457	0.8671	0.8435	0.9498
XGBoost	0.9764	0.6696	0.7852	0.7320
ANN algorithm	0.8000	0.7500	0.8000	0.9222
FedIFD	0.9700	0.9700	0.9700	0.8362
FL-BERT	0.8400	0.8400	0.8500	0.7597
MIC-XGB	0.9798	0.9839	0.9688	0.9763
VMD-ELM	0.9000	0.8945	0.9347	0.9043
MTMCN	0.9830	0.9364	0.9135	0.8968
Extra trees	0.9400	0.9089	0.8910	0.9122
MIFR-DFDIA	0.9904	0.9904	0.9905	0.9904

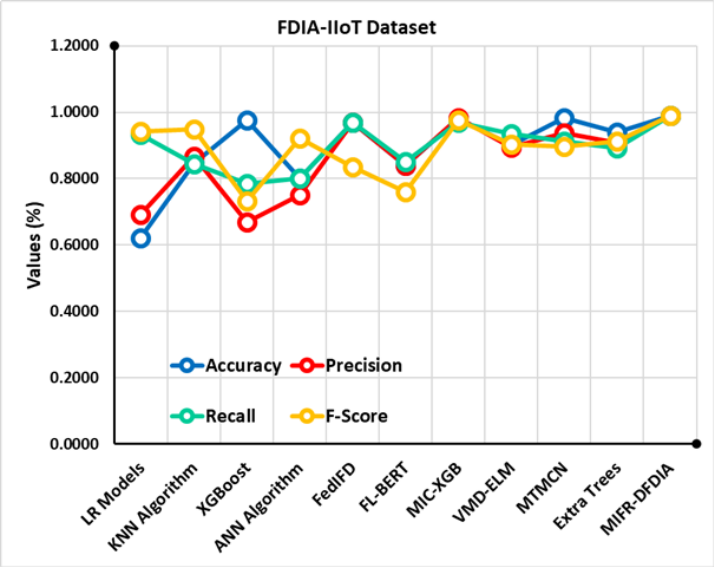

Figure 9. Comparative analysis of the MIFR-DFDIA model for the FDIA-IIoT dataset.

An ablation study inspects how parts of a model impact its performance by testing the system with particular components modified or removed. This method is represented to show which elements are significant and which are impacted with a minimum level. With the purpose of comparing the outcomes, we gain useful data regarding the behavior of the model. These findings support higher designs and optimistic outcomes. Table 4 shows the ablation study of the MIFR-DFDIA model using the FDIA-IIoT dataset. The experimentational outcome values indicate that the Stacked VAE only (without feature selection), MI+Stacked VAE (without WGAN), WGAN only (without feature selection and VAE), WGAN+MI (without VAE), and Stacked VAE+WGAN (without feature selection) techniques gain lower outcomes under diverse aspects. Moreover, the proposed MIFR-DFDIA (hybrid model with feature selection) algorithm provides greater performance with $accur_y$ of 0.9904, $preci_n$ of 0.9904, $recal_l$ of 0.9905, and $F1_{score}$ of 0.9904.

Table 4. Ablation study of the MIFR-DFDIA method for the FDIA-IIoT dataset.

Techniques	$\boldsymbol{Accur_y}$ (*100%)	$\boldsymbol{Preci_n}$ (*100%)	$\boldsymbol{Recal_l}$ (*100%)	$\boldsymbol{F1_{score}}$ (*100%)
Stacked VAE only (without feature selection)	0.9603	0.9564	0.9597	0.9587
MI+Stacked VAE (without WGAN)	0.9661	0.9643	0.9654	0.9651
WGAN only (without feature selection and VAE)	0.9722	0.9695	0.9707	0.9707
WGAN+MI (without VAE)	0.9796	0.9772	0.9782	0.9782
Stacked VAE+WGAN (without feature selection)	0.9846	0.9846	0.9848	0.9841
Proposed MIFR-DFDIA (hybrid model with feature selection)	0.9904	0.9904	0.9905	0.9904

4.3. Results and analysis of the IIoT dataset

Figure 10 illustrates the correlation matrix created by the MIFR-DFDIA model for the IIoT dataset. The result specifies that the MIFR-DFDIA approach effectively recognizes all classes.

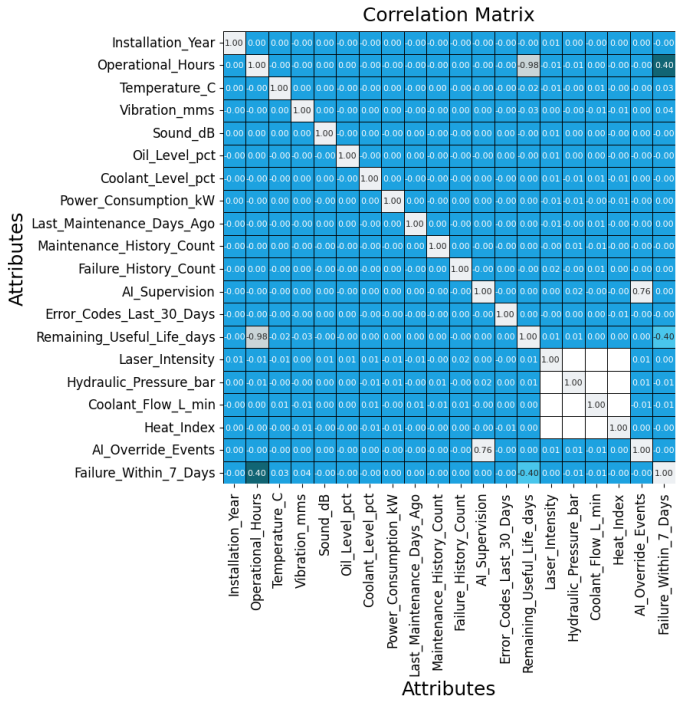

Figure 10. Correlation matrix of the MIFR-DFDIA system for the IIoT dataset.

Table 5 and Figure 11 represent an attack detection of the MIFR-DFDIA model on the IIoT dataset under different epochs. For epoch 500, the proposed MIFR-DFDIA method accomplishes an average $accur_y$ of 97.52%, $preci_n$ of 97.56%, $recal_l$ of 97.52%, and $F1_{score}$ of 97.52%. Similarly, for epoch 1500, the proposed MIFR-DFDIA approach gains an average $accur_y$ of 97.85%, $preci_n$ of 97.88%, $recal_l$ of 97.85%, and $F1_{score}$ of 97.85%. In addition, for epoch 3000, the proposed MIFR-DFDIA technique obtains an average $accur_y$ of 98.96%, $preci_n$ of 98.96%, $recal_l$ of 98.96%, and $F1_{score}$ of 98.96%.

Table 5. Attack detection of the MIFR-DFDIA model for the IIoT dataset.

Training cycle and overall	$Accur_y$		$Preci_n$		$Recal_l$		$F1_{score}$	
	Attack	Normal	Attack	Normal	Attack	Normal	Attack	Normal
500 Learning Epochs	99.02	96.02	96.14	98.99	99.02	96.02	97.56	97.48
Average	97.52		97.56		97.52		97.52	
Epoch-1000	99.01	96.46	96.55	98.98	99.01	96.46	97.76	97.71
Average	97.73		97.77		97.73		97.73	
Epoch-1500	99.05	96.65	96.73	99.03	99.05	96.65	97.88	97.82
Average	97.85		97.88		97.85		97.85	
Epoch-2000	99.05	96.8	96.87	99.03	99.05	96.8	97.95	97.9
Average	97.92		97.95		97.92		97.92	
Epoch-2500	99.01	97.26	97.31	98.99	99.01	97.26	98.15	98.12
Average	98.13		98.15		98.13		98.13	
Epoch-3000	99.36	98.56	98.57	99.35	99.36	98.56	98.96	98.96
Average	98.96		98.96		98.96		98.96	

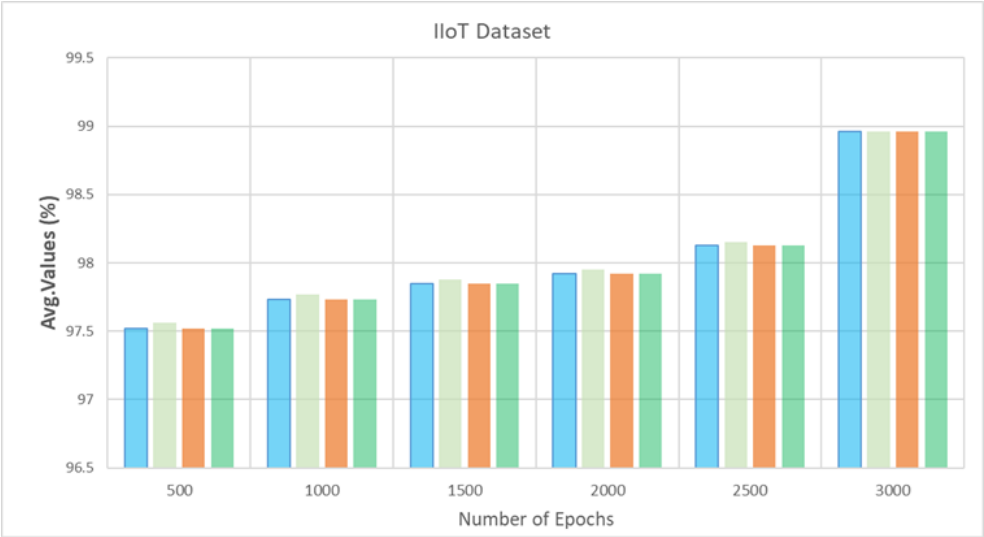

Figure 11. Average values of the MIFR-DFDIA model for the IIoT dataset.

Figure 12 displays the TRAN and VALD accuracy of MIFR-DFDIA technology for the IIoT dataset over 3000 epochs. Both curves progressively surge and deliberately converge, which specifies that the approach is learning effectively. The VALD accuracy continually remains slightly better than the TRAN accuracy, which infers that the method is not over-fitting and generalizes better to unnoticed data. The variations in accuracy are expected due to the task difficulty, but the overall upward trend indicates robust model performance.

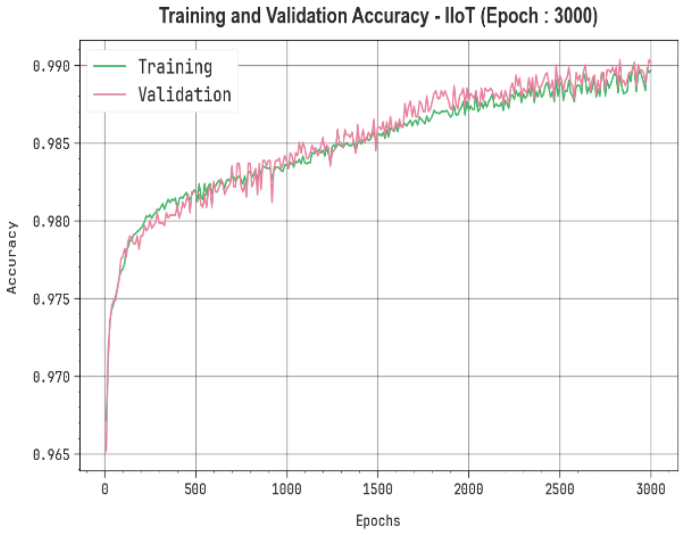

Figure 12. Accuracy curve of the MIFR-DFDIA algorithm for the IIoT dataset.

Figure 13 demonstrates the TRAN and VALD loss of the MIFR-DFDIA approach for the IIoT dataset over 3000 epochs. Both curves represent a consistent decline, signifying that the method efficiently reduces error during learning. The VALD loss stays slightly lesser than the training loss across most epochs. However, some instabilities are observed to achieve greater stability and consistency.

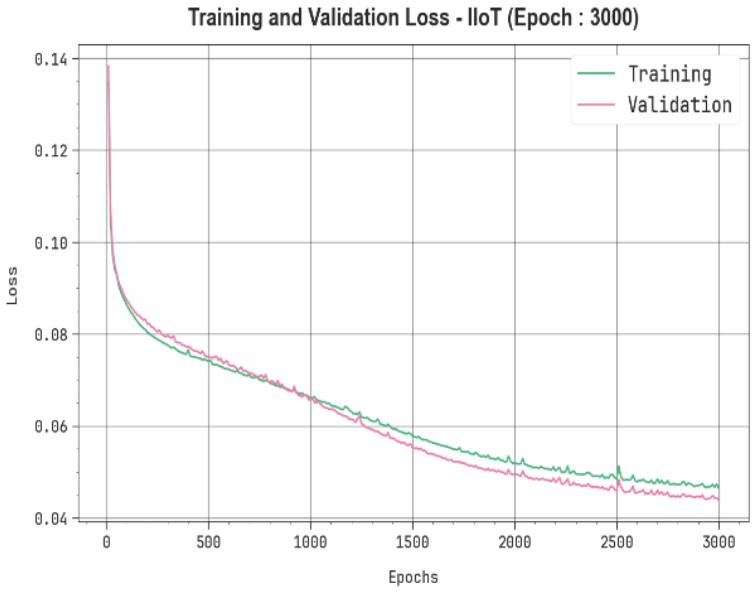

Figure 13. Loss curve of the MIFR-DFDIA algorithm for the IIoT dataset.

Table 6 and Figure 14 offer a comparison analysis of the MIFR-DFDIA method for the IIoT dataset with existing techniques. The results indicate that the models, including the SMA model, ARIMA method, transformer, Att-DNN, and decision tree, achieve lower perofrmances, while the Autoencoder and SVM models have slightly higher outcomes. Additionally, the BiGRU-LSTM and L2D2 (LSTM) methods obtain reasonable and closer results. Moreover, the proposed MIFR-DFDIA method obtains greater performance with $accur_y$ of 0.9896, $preci_n$ of 0.9896 $recal_l$ of 0.9896, and $F1_{score}$ of 0.9896. Thus, the presented MIFR-DFDIA model is highly effective for an attack detection process.

Table 6. Comparative analysis of the MIFR-DFDIA system for the IIoT dataset.

Methods	$\boldsymbol{Accur_y}$ (*100%)	$\boldsymbol{Preci_n}$ (*100%)	$\boldsymbol{Recal_l}$ (*100%)	$\boldsymbol{F1_{score}}$ (*100%)
BiGRU-LSTM	0.9832	0.8191	0.9666	0.8801
Att-DNN	0.8834	0.8665	0.8694	0.9073
L2D2 (LSTM)	0.9800	0.8541	0.7818	0.8201
SMA model	0.7820	0.9384	0.9062	0.8495
ARIMA method	0.8250	0.9331	0.8061	0.8441
Transformer	0.8930	0.9443	0.9708	0.8561
EMD-TF-BiLSTM	0.9250	0.9340	0.7929	0.9223
Decision tree	0.9000	0.9512	0.9460	0.9293
Autoencoder	0.9578	0.9556	0.9091	0.9450
SVM	0.9705	0.9178	0.9233	0.8977
MIFR-DFDIA	0.9896	0.9896	0.9896	0.9896

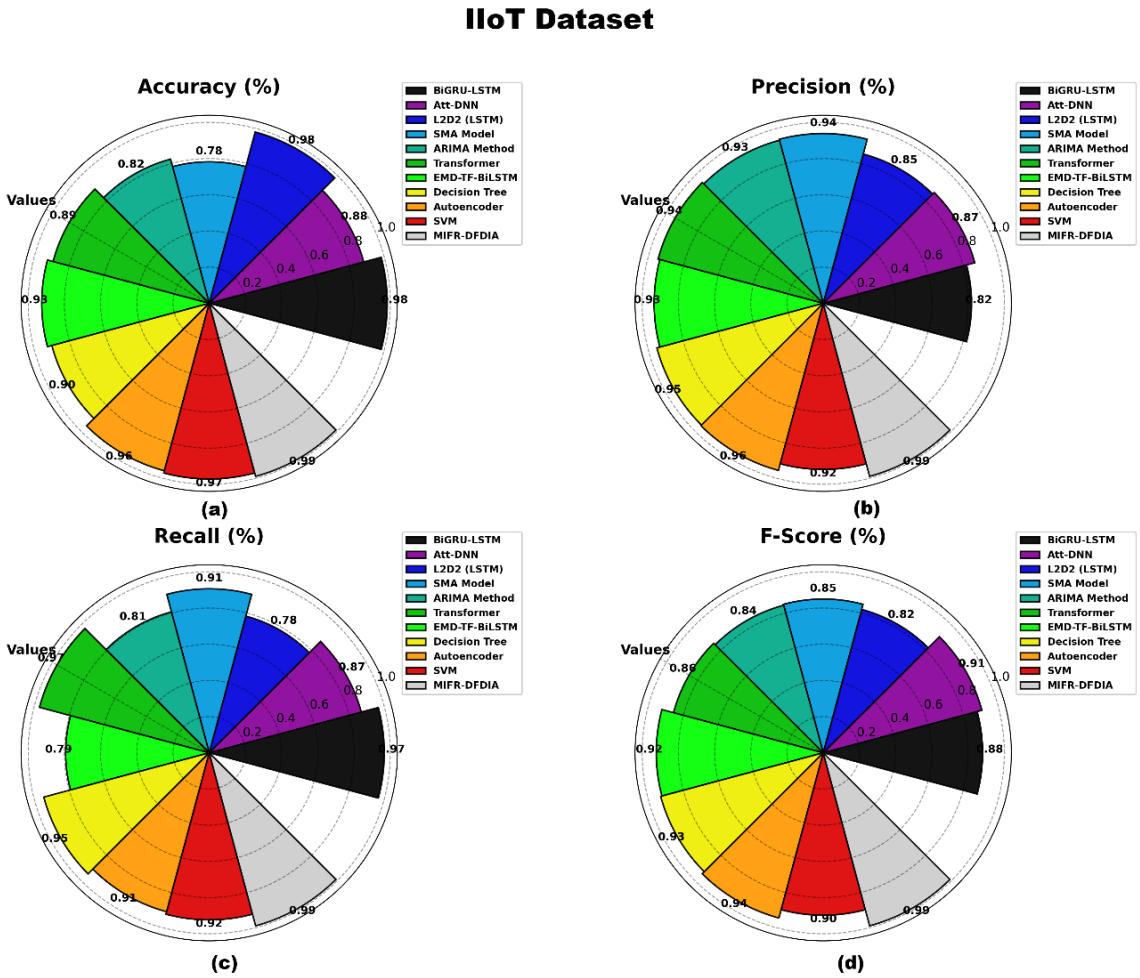

Figure 14. Comparative analysis of the MIFR-DFDIA system for the IIoT dataset.

An ablation study explores the contribution of elements within a technique by systematically eliminating or changing them and monitoring the impacts on performance. This method supports classifying the parts that are significant and those that have a slight influence. By examining these outcomes, researchers can better understand the system's inner workings and create more informed choices for enhancing design and efficacy. Table 7 represents the ablation study of the MIFR-DFDIAmodel using the IIoT dataset. The outcomes indicate that the Stacked VAE only (without feature selection), MI+Stacked VAE (without WGAN), WGAN only (without feature selection and VAE), WGAN+MI (without VAE), and Stacked VAE+WGAN (without feature selection) approaches obtain lesser outcomes under various measures. Moreover, the proposed MIFR-DFDIA (hybrid model with feature selection) model obtains a higher performance with $accur_y$ of 0.9896, $preci_n$ of 0.9896, $recal_l$ of 0.9896, and $F1_{score}$ of 0.9896.

Table 7. Ablation study of the MIFR-DFDIA approach for the IIoT dataset.

Methods	$Accur_y$ (*100%)	$Preci_n$ (*100%)	$Recal_l$ (*100%)	$F1_{score}$ (*100%)
Stacked VAE only (without feature selection)	0.9554	0.9564	0.9584	0.9569
MI+Stacked VAE (without WGAN)	0.9628	0.9632	0.964	0.9626
WGAN only (without feature selection and VAE)	0.9708	0.97	0.9707	0.9703
WGAN+MI (without VAE)	0.9775	0.9771	0.9763	0.9754
Stacked VAE+WGAN (without feature selection)	0.9831	0.9825	0.9831	0.9828
Proposed MIFR-DFDIA (hybrid model with feature selection)	0.9896	0.9896	0.9896	0.9896

4.4. XAI interpretation across two datasets

Figure 15 displays the importance of different features in a model's predictions using SHAP values under the FDIA-IIoT Dataset. The left side displays the mean absolute SHAP values, ranking attributes by their overall model contribution, where http_status_code and dst_port emerge as the most influential features. The right side visualizes the distribution of SHAP values for each feature across all samples, signifying how individual feature values push the model output toward higher or lower predictions. Features with greater SHAP values reliably push predictions in one direction, which helps in interpreting the model's decision-making process.

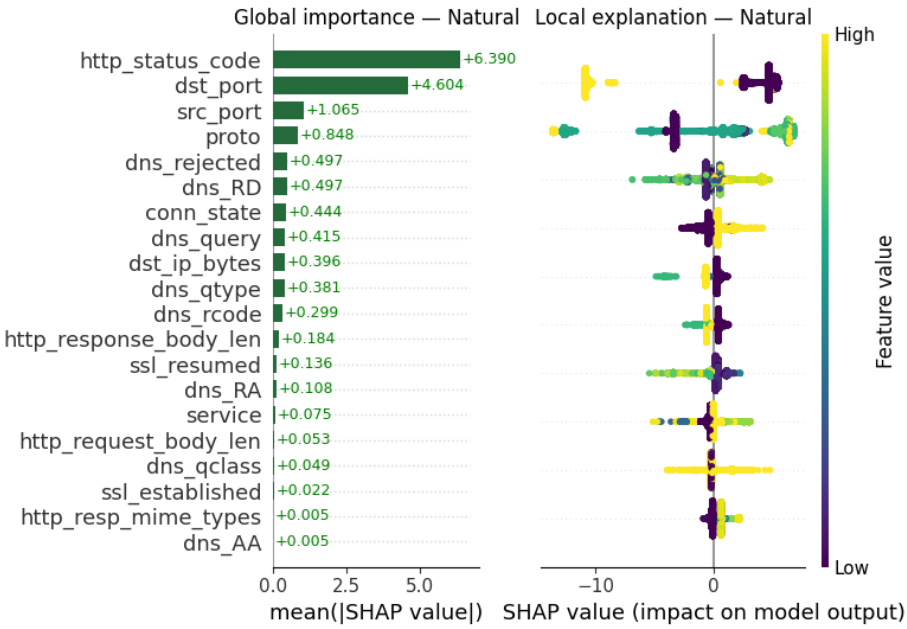

Figure 15. SHAP value of global and local explanation for the FDIA-IIoT dataset.

Figure 16 exemplifies feature importance for predicting a condition labeled "True" using SHAP values under the IIoT Dataset. The left side denotes that Remaining_Useful_Life_days is by far the most influential feature, followed by Operational_Hours, while others have much smaller impacts. The right side illustrates the distribution of SHAP values for individual samples, signifying how variations

in feature values either increase or decrease the system's output. This helps explain which factors most affect the model's decision for the "True" class.

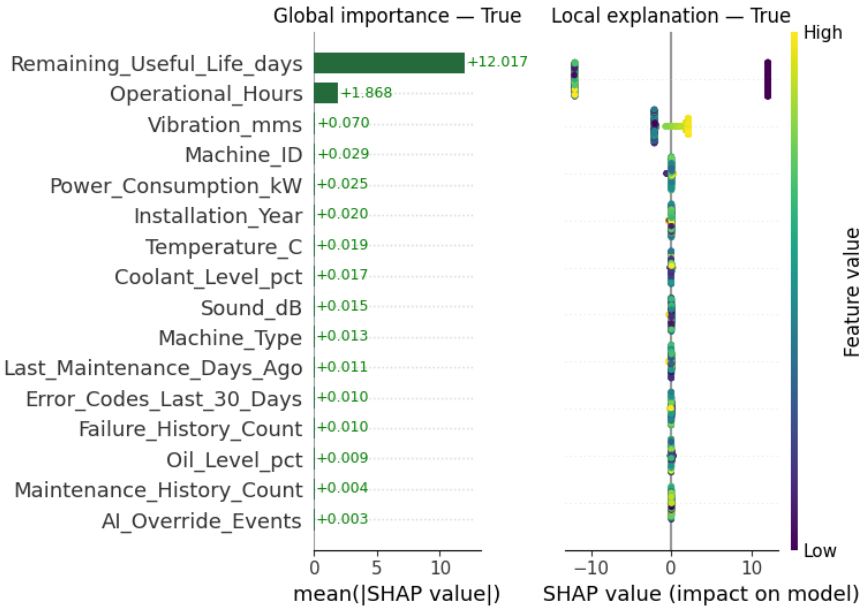

Figure 16. SHAP value of global and local explanations for the IIoT dataset.

The SHAP results (Figures 15 and 16) signify that features with higher contribution scores related to critical sensor measurements and network features are typically targeted during FDIA, such as abnormal signal deviations and communication inconsistencies. This alignment with domain knowledge ensures that the model learns meaningful attack patterns, while SHAP-based interpretability improves transparency, enabling operators to assess predictions and support reliable decision-making in AIoT systems.

5. Conclusions

In this study, a MIFR-DFDIA system is presented to improve cybersecurity resilience by precisely detecting and mitigating FDIA that compromise data integrity. Primarily, data preprocessing is applied to handle outliers, missing values, and feature standardization to guarantee high-quality inputs for analysis. MI is further employed for optimum dimensionality reduction to obtain the most significant features effectually. For classification, a hybrid approach such as SVAE and WGAN is deployed for efficient and precise detection of FDIA in cybersecurity distributed systems. Last, the XAI technique, SHAP, is leveraged to interpret model predictions and enhance transparency in decision-making. The experimental result analysis of the MIFR-DFDIA method is conducted against a benchmark dataset, and the comparative analysis reveals the supremacy of the MIFR-DFDIA method compared to other approaches in diverse metrics. In practical AIoT environments, adversaries may adjust by perturbing features recognized as vital through mutual information or SHAP analysis. To reduce such risks, future extensions may integrate randomized feature selection, adversarial training, and robustness-aware explainability to improve resilience against adaptive FDIA attacks.

Author contributions

M. A. AlAqil and H. K. Alkahtani: Conceptualisation; M. A. AlAqil, N. Allheeib and J. Khan: Data curation, Formal analysis; M. A. AlAqil, H. K. Alkahtani and N. Allheeib: Funding support, Investigation, Methodology; S. Almutairi: Project administration, Resources; M. A. AlAqil, H. K. Alkahtani, N. Allheeib, J. Khan, H. Alkhudhayr, S. M. Alenezi, M. B. Alharbi and S. Almutairi: Discussion, Writing–review and editing; S. M. Alenezi and M. B. Alharbi: Software, Supervision, Validation, Visualisation; H. K. Alkahtani; N. Allheeb; M. B. Alharbi, J. Khan and S. M. Alenezi: Writing–original draft. All authors have read and agreed to the published version of the manuscript.

Use of Generative-AI tools declaration

The authors declare they have used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

The authors extend their appreciation to the Deanship of Research and Graduate Studies at King Khalid University for funding this work through Large Research Project under grant number RGP2/271/46. This work was supported by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia, under Project Grant KFU261827; Ongoing Research Funding program (ORF-2026-609), King Saud University, Riyadh, Saudi Arabia; Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2026R384), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. The authors would like to thank the Deanship of Scientific Research at Shaqra University for supporting this work. The authors extend their appreciation to the Deanship of Scientific Research at Northern Border University, Arar, KSA for funding this research work through the project number "NBU-FPEJ-2026-1182-01".

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this manuscript.

References

1. F. Charmet, H. C. Tanuwidjaja, S. Ayoubi, P. F. Gimenez, Y. F. Han, H. Jmila, et al., Explainable artificial intelligence for cybersecurity: a literature survey, *Ann. Telecomm.*, **77** (2022), 789–812. https://doi.org/10.1007/s12243-022-00926-7
2. A. Shees, M. Tariq, A. I. Sarwat, Cybersecurity in smart grids: detecting false data injection attacks utilizing supervised machine learning techniques, *Energies*, **17** (2024), 5870. https://doi.org/10.3390/en17235870
3. F. Almutairy, L. Scekic, R. Elmoudi, S. Wshah, Accurate detection of false data injection attacks in renewable power systems using deep learning, *IEEE Access*, **9** (2021), 135774–135789. https://doi.org/10.1109/ACCESS.2021.3117230

4. A. Sayghe, Y. D. Hu, I. Zografopoulos, X. R. Liu, R. G. Dutta, Y. Jin, et al., Survey of machine learning methods for detecting false data injection attacks in power systems, *IET Smart Grid*, **3** (2020), 581–595. https://doi.org/10.1049/iet-stg.2020.0015
5. Y. Zhang, J. H. Wang, B. Chen, Detecting false data injection attacks in smart grids: a semi-supervised deep learning approach, *IEEE Trans. Smart Grid*, **12** (2021), 623–634. https://doi.org/10.1109/TSG.2020.3010510
6. Y. C. Ding, K. Ma, T. J. Pu, X. Y. Wang, R. Li, D. X. Zhang, A deep learning-based classification scheme for false data injection attack detection in the power system, *Electronics*, **10** (2021), 1459. https://doi.org/10.3390/electronics10121459
7. X. Y. Niu, J. N. Li, J. Y. Sun, K. Tomsovic, Dynamic detection of false data injection attack in smart grid using deep learning. In: *2019 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, 2019, 1–6. https://doi.org/10.1109/ISGT.2019.8791598
8. G. R. Mode, P. Calyam, K. A. Hoque, Impact of false data injection attacks on deep learning enabled predictive analytics, In: *NOMS 2020-2020 IEEE/IFIP Network Operations and Management Symposium*, Budapest, Hungary, 2020, 1–7. https://doi.org/10.1109/NOMS47738.2020.9110395
9. M. Elnour, M. Noorizadeh, M. Shakerpour, N. Meskin, K. Khan, R. Jain, A machine learning based framework for real-time detection and mitigation of sensor false data injection cyber-physical attacks in industrial control systems, *IEEE Access*, **11** (2023), 86977–86998. https://doi.org/10.1109/ACCESS.2023.3303015
10. D. Mukherjee, S. Chakraborty, A. Y. Abdelaziz, A. El-Shahat, Deep learning-based identification of false data injection attacks on modern smart grids, *Energy Reports*, **8** (2022), 919–930. https://doi.org/10.1016/j.egyr.2022.10.270
11. H. Moayyed, M. Mohammadpourfard, C. Konstantinou, A. Moradzadeh, B. Mohammadi-Ivatloo, A. P. Aguiar, Image processing based approach for false data injection attacks detection in power systems, *IEEE Access*, **10** (2022), 12412–12420. https://doi.org/10.1109/ACCESS.2021.3131506
12. D. B. Unsal, T. S. Ustun, S. M. S. Hussain, A. Onen, Enhancing cybersecurity in smart grids: false data injection and its mitigation, *Energies*, **14** (2021), 2657. https://doi.org/10.3390/en14092657
13. A. Sargolzaei, K. Yazdani, A. Abbaspour, C. D. Crane III, W. E. Dixon, Detection and mitigation of false data injection attacks in networked control systems, *IEEE Trans. Ind. Inform.*, **16** (2020), 4281–4292. https://doi.org/10.1109/TII.2019.2952067
14. V. Soni, D. Mehta, A. Shah, S. Dhingani, R. Gupta, S. Tanwar, et al., Explainable machine leaning-based false data injection classification framework for Avs, In: *2025 IEEE International Conference on Communications Workshops (ICC Workshops)*, Montreal, QC, Canada, 2025, 1924–1929. https://doi.org/10.1109/ICCWorkshops67674.2025.11162467
15. P. Jain, A. Rathour, A. Sharma, G. S. Chhabra, Bridging explainability and security: an XAI-enhanced hybrid deep learning framework for IoT device identification and attack detection, *IEEE Access*, **13** (2025), 127368–127390. https://doi.org/10.1109/ACCESS.2025.3590159
16. J. P. Ji, Y. Liu, J. Chen, Z. W. Yao, M. D. Zhang, Y. Y. Gong, False data injection attack detection method based on deep learning with multi-scale feature fusion, *IEEE Access*, **12** (2024), 89262–89274. https://doi.org/10.1109/ACCESS.2024.3418883
17. H. T. Feng, Y. H. Han, K. K. Li, F. Y. Si, Q. Zhao, Locational detection of the false data injection attacks via semi-supervised multi-label adversarial network, *Int. J. Electr. Power Energy Syst.*, **155** (2024), 109682. https://doi.org/10.1016/j.ijepes.2023.109682

18. H. Alamro, K. Mahmood, S. S. Aljameel, A. Yafoz, R. Alsini, A. Mohamed, Modified red fox optimizer with deep learning enabled false data injection attack detection, *IEEE Access*, **11** (2023), 79256–79264. https://doi.org/10.1109/ACCESS.2023.3298056
19. R. Huang, Y. C. Li, X. Wang, Attention-aware deep reinforcement learning for detecting false data injection attacks in smart grids, *Int. J. Electr. Power Energy Syst.*, **147** (2023), 108815. https://doi.org/10.1016/j.ijepes.2022.108815
20. E. Naderi, A. Asrari, A deep learning framework to identify remedial action schemes against false data injection cyberattacks targeting smart power systems, *IEEE Trans. Ind. Inform.*, **20** (2024), 1208–1219. https://doi.org/10.1109/TII.2023.3272625
21. H. B. Guo, J. Sun, Z. H. Pang, Residual-based false data injection attacks against multi-sensor estimation systems, *IEEE/CAA J. Automat. Sinica*, **10** (2023), 1181–1191. https://doi.org/10.1109/JAS.2023.123441
22. H. Zhou, D. C. Ren, H. X. Xia, M. Y. Fan, X. Yang, H. Huang, AST-GNN: An attention-based spatio-temporal graph neural network for interaction-aware pedestrian trajectory prediction, *Neurocomputing*, **445** (2021), 298–308. https://doi.org/10.1016/j.neucom.2021.03.024
23. M. Y. Fan, X. Q. Zhang, J. Hu, N. N. Gu, D. C. Tao, Adaptive data structure regularized multiclass discriminative feature selection, *IEEE Trans. Neural Networks Learn. Syst.*, **33** (2022), 5859–5872. https://doi.org/10.1109/TNNLS.2021.3071603
24. L. Q. Yang, Y. Zhai, Z. J. Li, Deep learning for online AC false data injection attack detection in smart grids: an approach using LSTM-autoencoder, *J. Network Comput. Appl.*, **193** (2021), 103178. https://doi.org/10.1016/j.jnca.2021.103178
25. J. W. Tian, C. Shen, C. H. Lin, M. Zhang, X. F. Xia, C. Ren, ADMM-based adversarial false data injection attacks against multi-label locational detection, *IEEE Trans. Depend. Secure Comput.*, **23** (2026), 263–277. https://doi.org/10.1109/TDSC.2025.3605689
26. J. W. Tian, C. Shen, B. H. Wang, X. F. Xia, M. Zhang, C. H. Lin, LESSON: Multi-label adversarial false data injection attack for deep learning locational detection, *IEEE Trans. Depend. Secure Comput.*, **21** (2024), 4418–4432. https://doi.org/10.1109/TDSC.2024.3353302
27. M. T. Nguyễn, T. L. Nhat, Identifying risk factors and survival prediction of heart failure patients using machine learning, *J. Sci. Technol. Inform. Commun.*, **1** (2025), 3–10.
28. C. Y. Liu, X. M. Yu, P. L. Wu, H. Y. Wang, A novel 3D indoor localization method integrating deep spatial feature augmentation and attention-based denoising, *Sci. Rep.*, **15** (2025), 33025. https://doi.org/10.1038/s41598-025-18549-y
29. P. N. Srinivasu, M. Sailaja, S. C. Narahari, P. Barsocchi, A. K. Bhoi, XAI driven software defect prediction using adaptive feature engineering coupled with autoencoder and multi-layer perceptron: an empirical study, *IEEE Access*, **13** (2025), 168693–168710. https://doi.org/10.1109/ACCESS.2025.3603451
30. False data injection attack dataset for industrial internet of things, 2025. Available from: https://zenodo.org/records/14864902.
31. Industrial IoT dataset (Synthetic). Available from: https://www.kaggle.com/datasets/canozensoy/industrial-iot-dataset-synthetic/data?select=factory_sensor_simulator_2040.csv.
32. H. Wang, J. Y. Yang, J. Sun, Z. Wang, Q. Z. Liu, S. X. Luo, FedIFD: Identifying false data injection attacks in internet of vehicles based on federated learning, *Big Data Cogn. Comput.*, **9** (2025), 246. https://doi.org/10.3390/bdcc9100246

33. T. Li, T. Xia, H. M. Zhang, D. Y. Liu, H. Zhao, Z. L. Liu, False data injection attacks detection based on stacking and MIC-DCXGB, *Sustainability*, **16** (2024), 9692. https://doi.org/10.3390/su16229692
34. A. Gueriani, H. Kheddar, A. C. Mazari, M. C. Ghanem, A robust cross-domain IDS using BiGRU-LSTM-attention for medical and industrial IoT security, *ICT Express*, 2025, In press. https://doi.org/10.1016/j.icte.2025.08.011
35. M. N. A. Ramadan, M. A. H. Ali, H. Jaber, M. Alkhedher, Blockchain-secured IoT-federated learning for industrial air pollution monitoring: a mechanistic approach to exposure prediction and environmental safety, *Ecotox. Environ. Safety*, **300** (2025), 118442. https://doi.org/10.1016/j.ecoenv.2025.118442

AIMS Press