Neural Computing and Applications (2023) 35:11459–11475
https://doi.org/10.1007/s00521-023-08319-0

ORIGINAL ARTICLE

An optimized model for network intrusion detection systems in industry 4.0 using XAI based Bi-LSTM framework

S. Sivamohan[1] · S. S. Sridhar[1]

Received: 23 May 2022 / Accepted: 16 January 2023 / Published online: 10 March 2023

Abstract
Industry 4.0 enable novel business cases, such as client-specific production, real-time monitoring of process condition and progress, independent decision making and remote maintenance, to name a few. However, they are more susceptible to a broad range of cyber threats because of limited resources and heterogeneous nature. Such risks cause financial and reputational damages for businesses, well as the theft of sensitive information. The higher level of diversity in industrial network prevents the attackers from such attacks. Therefore, to efficiently detect the intrusions, a novel intrusion detection system known as Bidirectional Long Short-Term Memory based Explainable Artificial Intelligence framework (BiLSTM-XAI) is developed. Initially, the preprocessing task using data cleaning and normalization is performed to enhance the data quality for detecting network intrusions. Subsequently, the significant features are selected from the databases using the Krill herd optimization (KHO) algorithm. The proposed BiLSTM-XAI approach provides better security and privacy inside the industry networking system by detecting intrusions very precisely. In this, we utilized SHAP and LIME explainable AI algorithms to improve interpretation of prediction results. The experimental setup is made by MATLAB 2016 software using Honeypot and NSL-KDD datasets as input. The analysis result reveals that the proposed method achieves superior performance in detecting intrusions with a classification accuracy of 98.2%.

Keywords Industry 4.0 · Cyber security · Krill herd optimization algorithm · Explainable artificial intelligence · Bidirectional long short-term memory

1 Introduction

The fourth industrial revolution (Industry 4.0) has been recently one of the key areas of discussion and research in the fields of engineering and management by academia and industry [1]. Digitalization has been recognized as a feasible technique to address the difficulties with the rise of industry 4.0 [2]. The concept of Industry 4.0 is built on developing technologies like cloud computing (CC), deep learning, fog computing, artificial intelligence (AI), and so on. Agriculture 4.0 is the fourth agricultural revolution that is expected to be accelerated by Industry 4.0 [3]. Techniques for Big Data Processing and algorithms are also widely utilized, to enhance security, efficiency, and system scalability. By exploiting virtual resources, Cloud technology establishes lowering costs, enhancing scalability and Cloud Manufacturing. To enable the sophisticated smart home system these technologies will have an impact not only in the manufacturing sector but also in day-to-day life by transforming traditional devices into smart products. Furthermore, the technologies have resulted in the rise of new business models like "multi-sided digital platforms", which are firms capable of linking two or more groups of people through a digital platform [4]. The primary goal of developing an Intrusion Detection System (IDS) is to detect various kinds of suspicious network behavior that a standard firewall cannot detect. Sensors, a detection engine, and a console are common components of an intrusion detection system. The various forms of IDS are utilized in several industrial applications [5]. There are two types of security-based solutions such as strategies for prevention and procedures for detection. Authentication and encryption are utilized in strategies based on prevention to protect

✉ S. Sivamohan
ss3983@srmist.edu.in

[1] Department of Computing Technologies, SRM Institute of Science & Technology, Kattankulathur, India

against possible attacks [6]. When preventative approaches fail the methods based on detection are employed. Signature detection is one of the detecting methods. For identifying malicious traffic, this approach employs well-known patterns known as Detecting anomalies. To detect anomalous activity in the system by hybrid methods these detection approaches have intended that merge one/more methods. A passive Intrusion Detection System is one in which a security breach is detected by an Intrusion Detection System sensor, the event is logged, and the console is notified [7].

Also, the industrial production environment's security landscape is developing fast [8]. To begin with, linking historically separated industries to the Internet implies that the "air gap" idea no longer protects them. Cloud services and wireless communications are rapidly being used in Cyber-Physical Production Systems (CPPSs) to connect various stakeholders along the supply chain, thus expanding the attack surface [9]. Attackers use the vulnerabilities to access the system and to have more serious effects. Therefore, for protecting the network perimeter effective IDS is established which is critical to interact freely without imposing excessive access controls while enabling various subsystems [10]. Numerous machine learning and deep learning techniques were developed in the recent years however, they are more susceptible to a broad range of cyber threats because of limited resources and heterogeneous nature. Such risks cause financial and reputational damages for businesses, well as the theft of sensitive information. To efficiently detect the intrusions, a novel intrusion detection system known as Bidirectional Long Short-Term Memory based Explainable Artificial Intelligence framework (BiLSTM-XAI) is developed in this paper. The prime contribution of this work is described as follows.

- The proposed BiLSTM-XAI approach provides better security and privacy inside the industry networking system by detecting intrusions with explanations.
- The significant features are selected from the databases using the Krill herd optimization (KHO) algorithm.
- Comparing the performance of the proposed method with the existing method in terms of different metrics to determine the effectiveness of the approach.

The rest of this paper is organized as follows: Sect. 2 illustrates the related literary works introduced by different authors based on intrusion detection systems. Section 3 depicts the proposed methodology consisting of different portions such as data pre-processing, feature selection and classification. Section 4 portrays the results and their discussions and finally, Sect. 5 concludes the paper.

2 Related works

Alohali et al. [11] introduced an artificial intelligence-enabled multimodal fusion-based intrusion detection system (AIMMF-IDS) for CCPS (cognitive cyber-physical system) in the environment industry 4.0. The CCPS has various limitations and security issues due to its challenging design. The artificial intelligence model was established for addressing the cyber security issues in the environment of industry 4.0. The results showed the detection performance was enhanced and the local optima problem was removed. Tahir et al. [12] proposed an experience-driven attack design with federated learning-based intrusion detection (EDADFL-ID) in industry 4.0. The false data injection attack (FDIA) detection algorithm was employed in industry 4.0 but it was broken the data privacy and the performance effectiveness of the distributed and dynamic environments were destroyed. To solve the issues, the EDADFL-ID method was introduced. The result showed the distributed environment has very high detection accuracy but suffers with security related problems.

The fast anomaly identification-based multi-aspect data streams (MDS-AD) for intelligence intrusion detection in industry 4.0 was established by Qi et al. [13]. The novel anomaly detection method has the combination of PCA, LSH (locality-sensitive hashing) and isolation forest technique. The proposed method followed some properties to achieve good results such as (a) the multi-aspect data was operated by LSH (b) the proposed method predicted the group anomalies from the results (c) the dimensionality was reduced by PCA. The experiments were conducted in the UNSW-NB15 dataset and it showed the best results. The drawback of this technique was information leakage while performing data reduction process.

Yang et al. [14] explained the stacked one-class broad learning system (ST-OCBLS) method for intrusion detection in industry 4.0. The ST-OCBLS method was used to efficiently keep the advantages of the training process. The network traffic data's hidden features were learned by using the decoding and encoding mechanism of the ST-OCBLS method. The experiments were conducted through various real-world intrusion detection tasks for obtaining the best performance and high-efficiency range when it faced difficulties in network data.

Ibitoye O et al. [15] developed a deep learning method for the detection of security threats in the case of IoT networks using python language. When comparing Self-normalizing Neural Networks (SNN) with Feed-forward Neural Networks (FNN), the accuracy rate of SNN IDS was 5% higher while using adversarial samples. A boT-IoT dataset is obtained for the analysis of the adversarial attack. Self-normalizing features are extracted from the

adversarial samples. Saghezchi et al. [16] expressed a machine learning approach for detecting the distributed denial-of-service (DDoS) attack in industry 4.0. The network traffic data were gathered from the real-world semiconductor production factory for predicting the limitations of an existing method. The traffic features in the dataset were extracted using Netmat tool. The 45 bidirectional network flow features were extracted and various labeled datasets were designed for the testing and training model of the machine learning method. The experimental results showed superior performance in accuracy and false-positive rate.

The novel collaborative learning-based intrusion detection system to predict intrusion in industry 4.0 was developed by Khoa et al. [17]. In collaborative learning, the filters were developed for preventing and detecting cyber-attacks. The data were collected with the help of filters in the network to train the cyber attack detection model-based deep learning algorithm. The trained model was shared with IoT gateways for improving the accuracy rate to detect the intrusion. The simulation was performed in real datasets and the result showed enhanced accuracy range and reduced network traffic during data exchanging process.

Li et.al [18] presented a new federated deep learning mechanism "DeepFed" to detect intrusions that are threatening industrial cyber physical systems (CPSs). Initially, by the utilization of convolutional neural network and gated recurrent unit (CNN-GRU) modules, the presence of cyber threats was detected. Subsequently, security, as well as privacy, was enhanced by adopting Paillier based secure communication protocol. The potential of DeepFed was examined by using a real industrial CPS dataset and the results revealed their greater performance.

Chowdhury et al. [19] discussed the AI-assisted web applications for COVID-19 vaccine prioritization. The XGBoost and random forest classifiers were utilized for training the designs and predicting the risk factors. The optimal vaccine distribution was finished by utilizing the predicted risk class. The deep learning algorithms were employed for enhancing the efficiency and accuracy of the prototype. But it had a limitation with high cost.

Krishnaveni et al. [20] discussed network intrusion detection for cloud computing based on feature selection and ensemble classification methods. Honeypot, NSL, and Kyoto were the datasets utilized for performing the statistical analysis. Experimentation results demonstrated that the scheme honeypot dataset was more effective and better when compared to other approaches. On the other hand, the detection rate was low.

Krishnaveni et al. [21] presented the effective feature classification and selection by ensemble approach for network intrusion detection based on cloud computing. The univariate ensemble feature selection method was utilized for selecting the valuable minimized feature sets with the intrusion datasets. The objective of the ensemble approach was to enhance the prediction accuracy. Experimentation results showed that the scheme acquired high accuracy and robustness with different classification tasks. But, the ensemble approach is failed to replace the deep neural network design.

Barnard et al. [22] discussed robust network intrusion detection by utilizing explainable artificial intelligence. The extreme gradient boosting (XGBoost) design was carried out for performing the supervised intrusion detections. The NSL-KDD database was utilized for evaluating the performance. Experimentation results showed that the scheme outperformed in detecting the attacks when compared to other approaches. But, it failed for supporting the multiclass classification.

Liu et al. [23] presented the framework to enhance the AI for detecting the intrusion by utilizing the data cleaning methods. The data cleaning and Explainable artificial intelligence (XAI) techniques were employed for enhancing the understandability and explainability of the intrusion detection alerts. The results demonstrated that the usage of AI explainability and data-cleaning methods provided quality explanations for the analysts. On the other hand, the algorithm's performances were not provided through the AIX360 toolkit. Table 1 presents the summary of the contributions of the existing method.

Motivated by the above mentioned state of art techniques, this paper intends to present a novel intrusion detection mechanism using BiLSTM-XAI approach which provides better security and privacy inside the industry networking system.

3 Proposed methodology

This section provides a detailed description of the proposed Bidirectional Long Short-Term Memory based Explainable Artificial Intelligence (BiLSTM-XAI) framework for accurate detection of industrial network intrusions. The intrusion detection system is comprised of three phases, namely pre-processing, feature selection and classification. In pre-processing, the data are cleaned and normalized and thus increasing the quality of data. Then the significant data features are extracted by means of feature selection process using the krill herd optimization (KHO) algorithm. Finally, the classification is carried out using the BiLSTM-XAI approach which classifies the data and detects the presence of intrusions accurately. These procedures are described elaborately in the following sub-sections. Figure 1 portrays the block diagram of the proposed intrusion detection system.

Table 1 Summary of existing works

Authors, Year and Refs no.	Technique	Objective	Advantages	Disadvantages
Alohali et al. (2022) and [11]	AIMMF-IDS	Artificial intelligence model was established for addressing the cyber security issues	Detection performance was enhanced and the local optima problem was removed	High-fidelity simulations were hard to maintain as well as computationally costly to run
Tahir et al. (2021) and [12]	EDADFL-ID	To solve the issues, the EDADFL-ID method was introduced	Very high detection accuracy	Suffers from security related issues
Qi et al. (2021)and [13]	MDS-AD	To achieve good results	Dimensionality was reduced by PCA	High computation time
Yang et al. (2022) and [14]	ST-OCBLS	Obtaining the best performance and high-efficiency range	High-efficiency range	Noise has severely affected the system's effectiveness
Ibitoye et al. (2019) and [15]	FNN	For classifying intrusion attacks in IoT networks	High accuracy rate	the Bad packets were generated from software bugs
Saghezchi et al. (2022) and [16]	Bidirectional network	Predicting the limitations of an existing method	Superior performance in accuracy and false-positive rate	High loss rate
Khoa et al. (2020) and [17]	Collaborative Learning technique	For improving the accuracy rate to detect the intrusion	Network traffic was reduced	The accuracy of intrusion detection was influenced by less number of training data
Li et.al (2019) and [18]	CNN	To enhance privacy and security	Security, as well as privacy, was enhanced	incapability to deal with different domain industrial CPSs
Chowdhury et al. (2022) and [19]	Deep learning algorithms	Training the designs and predicting the risk factors	Risk factors were predicted	High cost
Krishnaveni et al. (2022) and [20]	Ensemble classification method	To perform the statistical analysis	Honeypot dataset was more effective	Low detection rate
Krishnaveni et al.(2021) and [21]	Ensemble approach	Enhancing the prediction accuracy	High accuracy	The ensemble approach is failed to replace the deep neural network design
Barnard et al. (2022) and [22]	Extreme gradient boosting (XGBoost) design	Performing the supervised intrusion detections	High detection accuracy	It failed to support the multiclass classification
Liu et al. (2021)and [23]	Explainable artificial intelligence (XAI)	Enhancing the understandability and explainability of the intrusion detection alerts	High efficiency	Poor system performances

3.1 Data pre-processing

In a particular dataset, the data ranges are modified in data preprocessing to enhance the information acquisition and its operation [24]. High contrast variation is found between the maximum and minimum range of the dataset. During this process, the normalization of data reduces the difficulties in an algorithm. Data normalization has greater power while applying the classification of algorithms in neural networks. If the neural network gains a back-propagation technique, then the neural network is fully efficient and it will accelerate the training speed because of input normalization.

3.1.1 Data cleaning

The data redundancies, noises, errors and unwanted observations present in the dataset are removed using data cleaning procedure. This process permits only the relevant data to further process.

3.1.2 Normalization function

Data scaling plays a major role in the normalization function, which has a minimum and maximum algorithm, this will transform the net data value between $[-1, 1]$ and $[0, 1]$. The below equation gives the normalization formula,

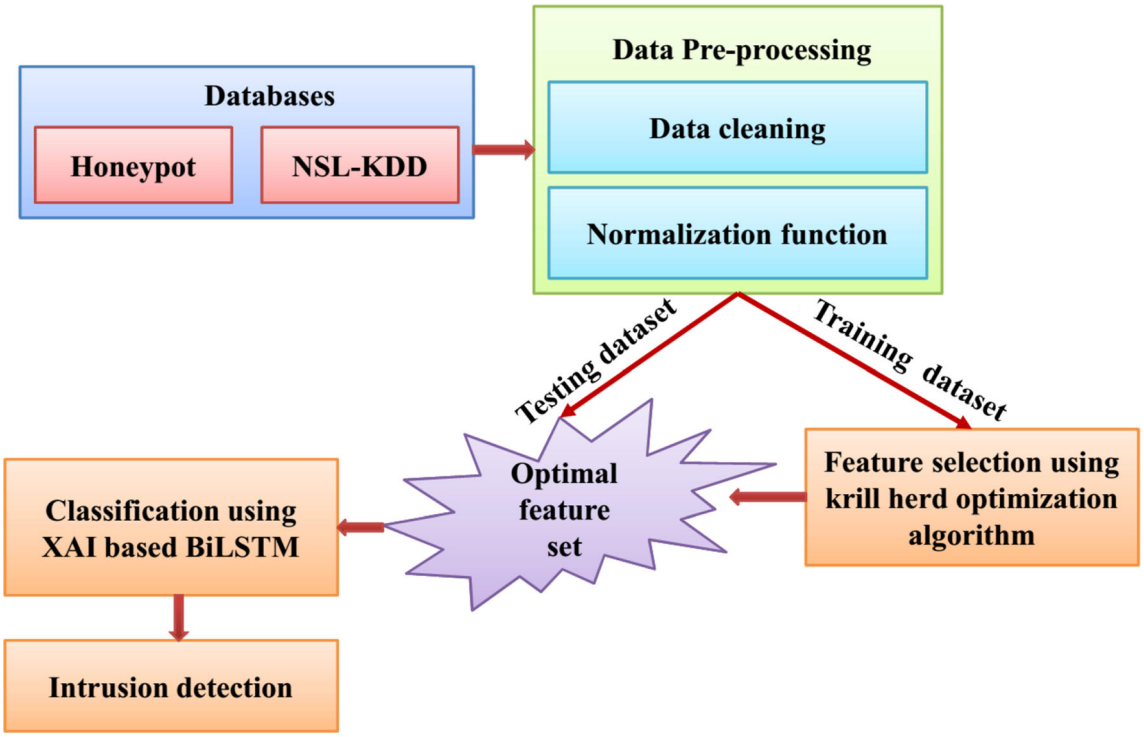

Fig. 1 Block diagram of the proposed intrusion detection system

$$I = \frac{d - d_{\text{MIN}}}{d_{\text{MAX}} - d_{\text{MIN}}} \tag{1}$$

From Eq. (1), the term I signifies the converted input value (i.e., normalized value). Also, the term d signifies actual value; d_{MAX} and d_{MIN} denote the maximum and minimum values of input variable d, respectively.

3.2 Feature selection process for IDS

As a frequent practice to select significant feature subsets, the feature selection process is gaining more attention [25]. Due to large dimensional and multiple featured data, the intrusion detection accuracy of model gets affected. The data features belonging to various classes contain different attributes and insignificant features that might cause the classifier to misclassify. The feature selection process minimizes data complexities by removing insignificant features from the dataset. It not only extracts redundant data but also reduces false positive rates in detecting intrusion. Moreover, it increases detection speed with a reduction in computation payload. Thus, the feature selection process creates more impact on detection accuracy and generalization capability of the model. Therefore, in this paper, we introduced the krill herd optimization (KHO) algorithm for feature selection which effectively selects the feature subsets by reducing data dimension and increasing detection accuracy and detection speed of intrusion detection system. A brief description of this KHO algorithm is modeled as follows.

3.2.1 Krill herd optimization (KHO) algorithm

The KHO algorithm simulated the krill behavior whereas every individual of the KH created their contribution with the process of moving [26]. The best solution is acquired when the krill individuals identify the food center. The KHO algorithm is according to the lagrangian as well as krill individual's evolution behavior with the capability to do exploitation and exploration in the optimization issues. In this KHO algorithm, the random value performs an important role. The time-dependent position of the individual's krill is calculated on the basis of three actions and they are; random diffusion, foraging activity and the movement induction of the krill individual. The KHO algorithm adopts the d-dimensional search space lagrangian design and it is expressed in the below equation;

$$\frac{\partial Z_j}{\partial t} = O\kappa_j + G\kappa_j + E\kappa_j \tag{2}$$

Here, the terms $O\kappa_j$, $G\kappa_j$ and $E\kappa_j$ indicates search agent's movement induction, foraging behavior and random diffusion, respectively. The optimization commences by initializing the parameters such as maximum diffusion speed E_{MAX}, krill position Z_j, maximal foraging speed v_g, maximal induced speed o^{MAX}, the maximal number of

iterations J_{MAX} and numbers of krill O. Compute the movement for every krill. The communal effects between krill individuals lead to movements and they attempt to conserve the higher density. σ_j denote the motion induction direction and which is evaluated with the local target and the density of the repulsive swarm. Then it is expressed as;

$$O\kappa_j^{\mathrm{NEW}} = o^{\mathrm{MAX}}\sigma_j + x_o O\kappa_j^{\mathrm{old}} \tag{3}$$

The inertia weight is represented by x_o, the last motion is denoted by $O\kappa_j^{\mathrm{OLD}}$ and maximal induced speed is represented as o^{MAX}.

$$\sigma_j = \sigma_j^{\mathrm{LO}} + \sigma_j^{\tau} \tag{4}$$

$$\sigma_j^{\mathrm{LO}} = \sum_{k=1}^{Oo} \hat{\kappa}_{j,k}\hat{Y}_{j,k} \tag{5}$$

$$\hat{Y}_{j,k} = \frac{Y_k - Y_j}{\left\|Y_k - Y_j\right\| + a} \tag{6}$$

$$\hat{\kappa}_{j,k} = \frac{\kappa_j - \kappa_k}{\kappa^{\mathrm{WO}} - \kappa^{\mathrm{BEST}}} \tag{7}$$

The worst and best krill individuals are represented by κ^{BEST} and κ^{WO}. The smaller positive number is represented by a. The numbers of neighbors are denoted by O_o. Also, the terms σ_j^{τ} and σ_j^{LO} denote target direction effect and local effect offered by neighbors, respectively; κ_j and κ_{k} depicts fitness function of jth krill and kth neighbor, respectively; Y_{k} and Y_j indicates the corresponding positions of jth krill and kth neighbor, respectively.

$$\sigma_j^{\tau} = D^{\mathrm{BEST}}\hat{\kappa}_{j,\mathrm{BEST}}\hat{Y}_{j,\mathrm{BEST}} \tag{8}$$

D^{BEST} represents the krill individual efficient individual through best fitness is expressed in the below equation;

$$D^{\mathrm{BEST}} = 2\left(\Re + \frac{J}{J_{\mathrm{MAX}}}\right) \tag{9}$$

$$E_{t,j} = \frac{1}{5O}\sum_{k=1}^{O} Y_j - Y_k \tag{10}$$

where $\Re$ depicts random value which lies in the range [0,1], J signifies current iteration, $E_{t,j}$ indicates sensing distance.

The foraging movement is defined as which is according to food location and previous experiences with the food location. Then it is expressed as;

$$G\kappa_j = v_g\eta_j + x_g G\kappa_j^{\mathrm{OLD}} \tag{11}$$

Inertia weight with foraging movement is represented by x_g, η_j signifies fitness value of jth krill and the krill best objective is represented by η_j^{BEST}.

$$\eta_j = \eta_j^{\mathrm{FOOD}} + \eta_j^{\mathrm{BEST}} \tag{12}$$

$$\eta_j^{\mathrm{FOOD}} = D^{\mathrm{FOOD}}\hat{\kappa}_{j,\mathrm{FOOD}}\hat{Y}_{j,\mathrm{FOOD}} \tag{13}$$

The food coefficients are represented by D^{FOOD} and it is expressed as;

$$D^{\mathrm{FOOD}} = 2\left(1 - \frac{J}{J_{\mathrm{MAX}}}\right) \tag{14}$$

The jth krill individual best objectives are determined by η_j^{BEST} and it is expressed in the below equation

$$\eta_j^{\mathrm{BEST}} = \hat{\kappa}_{j,j\mathrm{BEST}}\,\hat{Y}_{j,j\mathrm{BEST}} \tag{15}$$

The food centers for iterations are computed and it is expressed as;

$$Y^{\mathrm{FOOD}} = \frac{\sum_{j=1}^{O}\frac{1}{\kappa_j}Y_j}{\sum_{j=1}^{O}\frac{1}{\kappa_j}} \tag{16}$$

The physical diffusion movement is described on the basis of the diffusion speed for maximum and the random direction vectors are expressed in the below equation;

$$E\kappa_j = E^{\mathrm{MAX}}\left(1 - \frac{J}{J_{\mathrm{MAX}}}\right)\Phi \tag{17}$$

The random direction vector is represented by Φ which lies between -1 and 1.

To enhance the performance of KHO, the genetic reproduction mechanisms mutation and crossover are merged through KHO.

The below expression represents the crossover function of Y_j's nth component,

$$Y_{j,n} = \begin{cases} Y_{s,n}, & \Re_{j,n} < c_o \\ Y_{j,n}, & \text{else} \end{cases} \tag{18}$$

Then, crossover probability $c_o = 0.2\hat{\kappa}_{j,j\mathrm{BEST}}$

$$Y_{j,n} = \begin{cases} Y_{h\mathrm{BEST},n} + v\left(Y_{q,n} - Y_{r,n}\right) & \Re_{j,n} < \mathrm{Nu} \\ Y_{j,n} & \text{else} \end{cases} \tag{19}$$

The mutation probability is denoted by N_u and it is set to $N_u = 0.05/\hat{\kappa}_{j,j\mathrm{BEST}}$.

The krill's position vector in the interval $|t + \Delta t|$ is found using below expression,

$$Y_j(t + \Delta t) = Y_j(t) + \Delta t\frac{\partial Y_J}{\partial t} \tag{20}$$

3.3 Classification using BiLSTM-explainable artificial intelligence (BiLSTM-XAI)

In IDS, the classification task becomes an indispensable step, in which it classifies the input databases into two distinct categories by detecting whether the network is affected by any intrusions or not. To achieve efficient classification results, we developed a Bidirectional Long

Short-Term Memory based Explainable Artificial Intelligence framework (BiLSTM-XAI) which determines the presence of any unauthorized and abnormal behavior of the network accurately. The elaborated illustrations of these techniques are detailed in the following sub-sections.

3.3.1 Bidirectional long short-term memory (BiLSTM)

The information in Bidirectional Long-Short Term Memory (BiLSTM) is stored both forward and backward of the neural network [27]. The LSTM model is given an encoded sequence of Inception model characteristics. From sign language videos the temporal information/characteristics are extracted utilizing the LSTM models. LSTM cells comprise the LSTM model, which is utilized to discover long-range contextual links as well as to understand common temporal patterns in the input sequences from learned feature sequences.

$$j_p = \mu\left(Z_j \cdot \left[d_{p-1}, g_{p-1}, y_p\right] + a_j\right) \tag{21}$$

$$e_p = \mu\left(Z_e \cdot \left[d_{p-1}, g_{p-1}, y_p\right] + a_e\right) \tag{22}$$

$$d_p = e_p \cdot d_{p-1} + j_p \cdot \tilde{d}_p \tag{23}$$

$$q_p = \mu\left(Z_\circ \cdot \left[d_p, g_{p-1}, y_p\right] + a_\circ\right) \tag{24}$$

$$g_p = q_p \cdot \tanh\left(d_p\right) \tag{25}$$

y_p, g_p and d_p denote the input sequence, the output sequence, and the memory's state at any given time p. Also, the input gate, the forget gate, and output gate are denoted by j_p, e_p and q_p. The corresponding bias vectors of input gate, forget gate and output gate are denoted as a_j, a_e and $a_\circ$.The activations of the cells are depicted utilizing $\tilde{d}$. These values are the same size as the input vector. Nonlinear sigmoid functions are represented by the symbol μ. An LSTM layer made up of stacked LSTM cells can communicate and utilize similar weights as another layer. To create LSTM of the bidirectional/unidirectional these LSTM layers can be utilized. Here, in a BiLSTM the two layers work in opposite temporal directions. In the finding of long-term bidirectional relationships between time steps, these layers are utilized. Therefore, the output included features from both past time steps, and the future time step is one of the benefits of utilizing BiLSTM. Two bidirectional LSTM layers, each one with 256 stacked LSTM blocks are made of the BiLSTM model. To classify encoded sequences the softmax is employed after the BiLSTM layers. The Inception model after training, then feed the extracted features to the BiLSTM model and from the temporal sequences the features are extracted.

3.3.2 Explainable artificial intelligence (XAI)

The development of XAI techniques is to accomplish higher transparency and generate more explanations for AI system [28]. The XAI field of research is exploring numerous mechanisms that will permit the performance of autonomous intelligent systems to be understandable and interpretable to human beings. The interaction between humans and machines is bridging the gap between social science and data science that leads to advanced AI technology and also contributes toward accountable, transparent and reasonable AI. Generally, a detailed explanation paves the way to analyze the strengths and constraints of learning frameworks and thus facilitates trustworthy and understandability of the model. Post hoc explanation is a kind of explanation model which extracts the data based on black box model for making better decisions. It offers valuable data predominantly for both the end-users and practitioners by providing instant explanation instead of internal working. The main objective of XAI is to improve the transparency of learning techniques in making decisions regarding predicted output. In this paper, two post hoc explanation models such as SHAP and LIME are employed which are deliberated briefly as follows.

3.3.2.1 Local interpretable model-agnostic explanations (LIME) The LIME approach produces explanations for all the instances generated by the black box model. The explanation of the LIME model depends on the performance of classifier in the closeness of data instances to be described based on local surrogate model that is numerically illustrated in equation format as follows.

$$\text{Explanation}(y) = \text{ArgMin}_{G \in g} \int \left(F, G, \pi_y\right) + \Psi(G) \tag{26}$$

Here,y signifies the data instances to be explained, $\int \left(F, G, \pi_y\right)$ depicts fidelity term, $\Psi(G)$ represents complexity term, F denote black box model and G implies the explained data instance. The local surrogate model attempts to harmonize the data instance closeness to the predicted result. In the initial stage, the LIME model utilized perturbation concept to produce data features from the raw databases. Unlike this, the LIME model is implemented in different means by the consideration of univariate distributed features to determine the distribution of all features. Based on the frequency of data classes the sampling is made for categorical features while for numerical features three alternatives are followed. Initially, the raw database is assembled into bins depending on quantiles; a single bin is selected randomly from the group and sampled consistently between maximum and minimum of chosen bins. Next, LIME resembles the actual distribution of numerical features via normal distribution. Then,

by means of kernel density function, the original distribution of numerical features is approximated.

3.3.2.2 Shapley additive explanations (SHAP) For generating the explanations the Shapley additive explanations (SHAP) are utilized through the evaluation of a second human [2]. The SHAP gradient explainer and deep SHAP explainer combined the ideas with the integrated gradient. The kernel SHAP algorithm is selected for the current study. It is the model-agnostic approach and it is used for evaluating the SHAP values. SHAP kernel explainer operates for all designs but it is slowest than other design class-specific algorithms. It provides the best results and it also provided accurate SHAP values. The features contribution toward the prediction of model's output is described by each SHAP value. The SHAP makes the learning technique more explainable by visualizing the contribution of each feature. This method utilized the coalition's concept for computing the SHAP values.

$$\phi_k(y) = \frac{1}{N}\sum_{n=1}^{N}\phi_k^n \tag{27}$$

$$\phi_k^n = \hat{g}\left(y_{+k}^n\right) - \hat{g}\left(y_{-k}^n\right) \tag{28}$$

The term $\phi_k(y)$ describes the Shapley feature values of predicted y instances, ϕ_k^n signifies mean marginal contribution, $\hat{g}\left(y_{+k}^n\right)$ depicts black box prediction without replacement of kth feature and $\hat{g}\left(y_{-k}^n\right)$ implies black box prediction with replacement of kth feature.

3.3.3 Combined BiLSTM-XAI approach for efficient intrusion detection

Figure 2 describes the proposed BiLSTM-XAI approach to efficiently classify the intrusions present in the industrial network. The step by step procedure of BiLSTM-XAI approach is described as follows.

Step 1 The input databases, namely Honeypot and NSL-KDD, are injected into the BiLSTM framework that classifies the data features and detects the abnormal features if any present in the network.

Step 2 The extracted features of the BiLSTM framework may contain some inappropriate loss functions.

Step 3 Due to the loss functions, the detection accuracy gets diminished thereby causing misclassification results.

Step 4 Therefore, it necessitates interpretable explanations with justifications for the misclassified result to prevent the network from future attacks.

Step 5 This mechanism improves the transparency of the proposed intrusion detection system in making decisions regarding interpretation of predictions.

Step 6 To make this happen, this paper introduced the explainable AI models, namely LIME and SHAP models.

Step 7 The XAI approaches increase the interpretation efficiency by its ability to understand the impact of the malicious data.

Step 8 Thus, the BiLSTM-XAI approach determines the presence of any unauthorized and abnormal behavior (i.e., intrusions) of the network.

4 Result and discussion

In this section, the performance analysis based on the intrusion detection system is presented. The simulations are implemented in MATLAB 2016 (a) in a system containing an i3 processor and 4 GB RAM. The proposed method is examined by two datasets, namely Honeypot and NSL-KDD, respectively [29]. The datasets are partitioned into two sub-classes for training and testing with the proportion 70:30, respectively. The parameter settings of the algorithm are depicted in Table 2.

4.1 Performance metrics

Some of the performance metrics such as accuracy (A), precision (P), specificity (SP), F1-score, Recall (R) and Matthews correlation coefficient (MCC) are applied to evaluate the efficiency of the proposed method. Also, runtime and speedup time are computed. The mathematical expression for each metric is given below.

$$\text{Accuracy}\,(A) = \frac{t_{\text{pos}} + t_{\text{neg}}}{t_{\text{pos}} + f_{\text{pos}} + t_{\text{neg}} + f_{\text{neg}}} \tag{29}$$

$$\text{Precision}\,(P) = \frac{t_{\text{pos}}}{t_{\text{pos}} + f_{\text{pos}}} \tag{30}$$

$$\text{Specificity}\,(\text{SP}) = \frac{t_{\text{neg}}}{t_{\text{neg}} + f_{\text{pos}}} \tag{31}$$

$$\text{Recall}\,(R) = \frac{t_{\text{pos}}}{t_{\text{pos}} + f_{\text{neg}}} \tag{32}$$

$$F1-\text{score} = \frac{t_{\text{pos}}}{\frac{\left(f_{\text{pos}}+f_{\text{neg}}\right)}{2} + t_{\text{pos}}} \tag{33}$$

$$\text{MCC} = \frac{t_{\text{pos}} \times t_{\text{neg}} - f_{\text{pos}} \times f_{\text{neg}}}{\sqrt{\left(t_{\text{pos}} + f_{\text{neg}}\right)\left(t_{\text{pos}} + f_{\text{pos}}\right)\left(t_{\text{neg}} + f_{\text{pos}}\right)\left(t_{\text{neg}} + f_{\text{neg}}\right)}} \tag{34}$$

4.1.1 Receiver operating characteristic (ROC)

The ROC curve is formed by calculating the true positive value against the false positive value at a variety of thresholds.

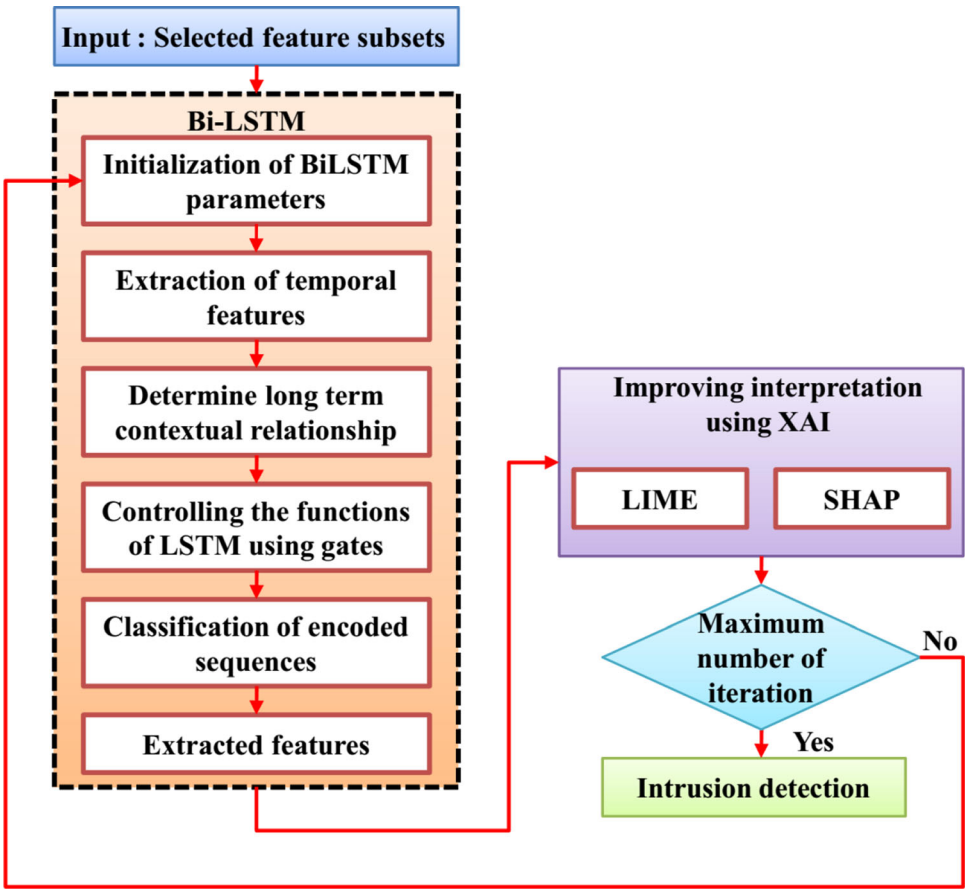

Fig. 2 Proposed BiLSTM-XAI approach

Table 2 Parameter settings

Parameters	Ranges
Size of the population	50
Maximum number of iterations	100
Diffusion speed	0.006
Foraging speed	0.2
Learning rate	0.001
Batch size	64
Dropout rate	0.5

The term $t_{pos}, t_{neg}, f_{pos}, f_{neg}$ in the equations represents true positives, true negatives, false positives and false negatives, respectively.

4.2 Dataset description

The dataset details are listed below in Table 3.

4.2.1 Dataset 1: honeypot

Five features like source_port_number, start_time, source_ip_address, destination_ip_address and source_port_number are dropped. In categorical features, one-hot encoding is applied so that 47 features were formed including labels and by using the Min–Max scaler, those extracted features are normalized. Without the label, a total of 45 features are available.

4.2.2 Dataset 2: NSL-KDD

By the utilization of Min–Max scalar, the numerical features are normalized and then via one hot encoding process, the categorical features are transformed into number

Table 3 Dataset description

Dataset	Number of features available	Number of features selected
Honeypot	47	45
NSL-KDD	124	122

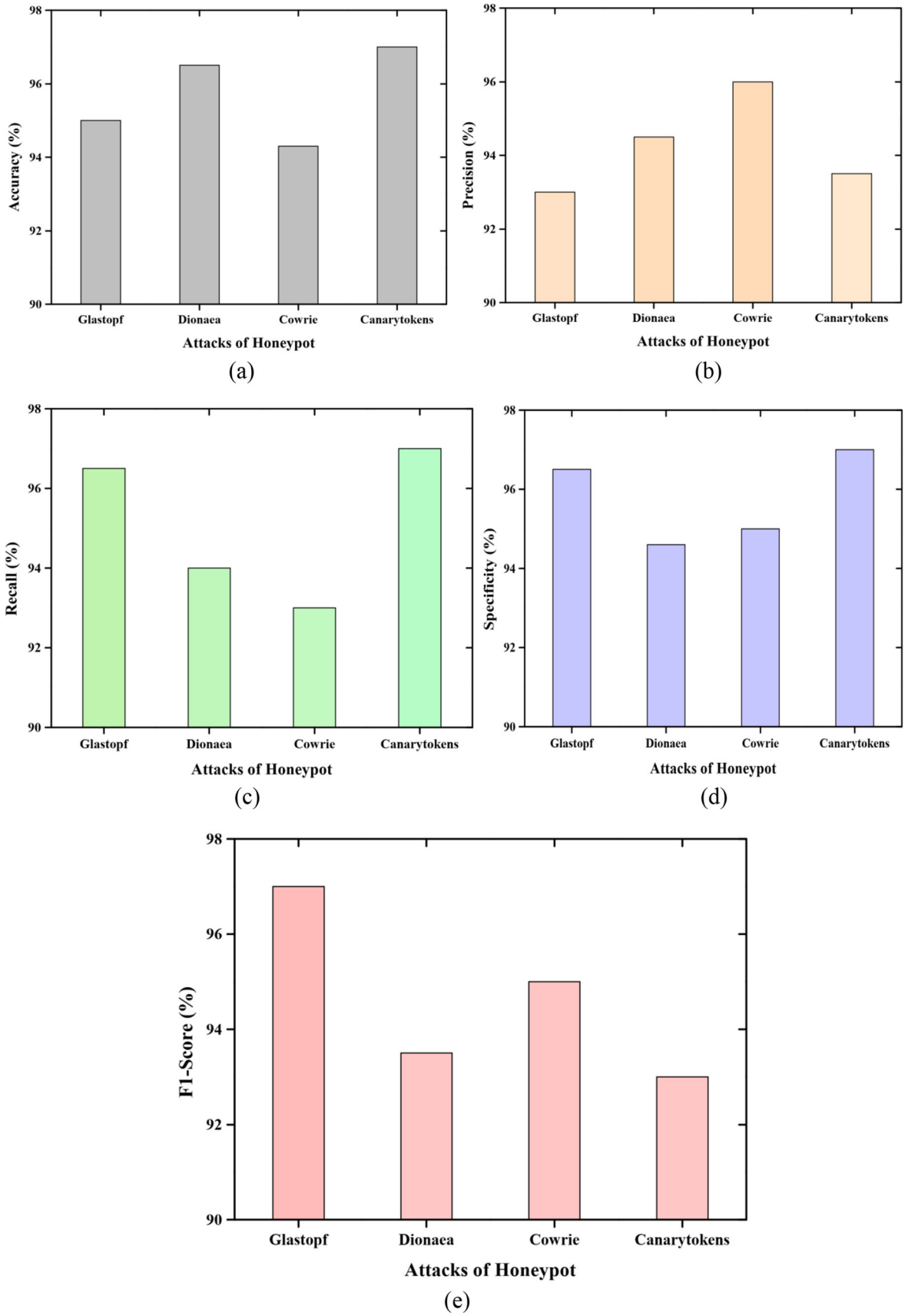

Fig. 3 Performance analysis of various attacks in Honeypot dataset. **a** Accuracy, **b** precision, **c** recall, **d** specificity, **e** F1-score

Table 4 Analysis of prediction performance with and without feature selection method

Accuracy (%)	
Classifier without KHO method	Classifier with KHO method
92.35%	98.2%

of dummy numerical values. This procedure makes 42 features of dataset to increase its number and thereby forming 124 features with labels and 122 features without labels.

4.3 Performance analysis

This section describes the performance analysis of two different datasets with respect to the different attacks of both datasets. Also, various performance metrics like accuracy, precision, recall, specificity, F1-score and ROC curve are graphically analyzed and discussed below. Figure 3 depicts the evaluation of the performance rate of dataset-1 Honeypot with various attacks. Different attacks like Glastopt, Dionaea, Cowrie and Canarytokens are analyzed. The graph is plotted between different attacks and performance rates. The accuracy, precision, recall, specificity, and F1-score of each attack are graphically represented. The analysis shows that the proposed method has obtained higher efficiency in detecting various attacks with a better performance rate. The accuracy rate of Glastopt, Dionaea, Cowrie and Canarytokens in the Honeypot dataset is 95.4, 96.8, 94 and 97.2%. Table 4 describes the result of classification performance with and without feature selection method.

The graphical representation of performance analysis of the NSL-KDD dataset with different attacks such as DoS, R2L, U2R and Probe are shown in Fig. 4. The accuracy, precision, recall, specificity, and F1-score of each attack are determined and analyzed. The result shows that the performance rate of all attacks has achieved better efficiency. The accuracy rate of DoS, R2L, U2R and Probe in the NSL-KDD dataset is 96, 95.6, 97.2 and 95.6%. The confusion matrix for attack detection is given in Fig. 5. Figure 5a shows the confusion matrix of Honeypot dataset, here the Glastopt attack has achieved higher accuracy of 93.24%. Figure 5b portrays the confusion matrix of NSL-KDD dataset, here the DoS attack has obtained higher performance accuracy of about 96.48%.

Figure 6 represents the analysis of the detection rate of Honeypot and NSL-KDD datasets. The graph shows that the Honeypot dataset has achieved a higher detection rate compared to the NSL-KDD dataset. The obtained detection rate of Honeypot is 97.2% and the NSL-KDD dataset is 95.8%, respectively.

The performance rate evaluation of various metrics like accuracy, specificity, precision, recall and F1-score are graphically represented in Fig. 7. The proposed BiLSTM-XAI method achieves greater performance rate than other compared methods. From the graph, the proposed BiLSTM-XAI method attains better accuracy of 98.4%, precision of 95.8%, specificity of 95.9%, recall of 97.2% and F1-score of 95.1% while other methods obtained slightly lesser values than the proposed method. Figure 8 illustrates the receiver operating characteristic (ROC) curve analysis of various methods. The proposed BiLSTM-XAI method achieves a higher ROC rate compared to DeepFed, AIMMF-IDS, EDADFL-ID and MDS-AD.

Figures 9a, b depict the comparative analysis of various methods in terms of different attacks in the Honeypot dataset and NSL-KDD dataset. The Honeypot dataset has four different attacks like Glastopt, Dionaea, Cowrie and Canarytokens whereas the NSL-KDD dataset comprises DoS, R2L, U2R and Probe attacks. The attack values of the proposed BiLSTM-XAI method are higher when compared to existing methods. From the comparative analysis, the performance of the proposed BiLSTM-XAI method is higher and the NSL-KDD dataset achieves better performance than dataset − 1.

4.4 Discussions

The comparative analysis of classification accuracy of various methods is shown in Fig. 10. The proposed BiLSTM-XAI method is compared with federated deep learning mechanism (DeepFed), artificial intelligence enabled multimodal fusion-based intrusion detection system(AIMMF-IDS), experience driven attack design with federated learning based intrusion detection (EDADFL-ID) and Multi-Aspect Data Streams-anomaly detection(MDS-AD). The classification accuracy of the proposed BiLSTM-XAI method is higher with compared to the existing methods. From the comparative analysis, the classification accuracy of the proposed BiLSTM-XAI method is about 98.2%. Figure 11 shows the Matthews correlation coefficient (MCC) analysis with respect to different methods. The proposed BiLSTM-XAI method achieves high MCC rate of about 0.96 while others produced less rate as compared to proposed method.

5 Conclusion

In this paper, a novel intrusion detection system is proposed to provide security and privacy inside the industry networking system. The proposed intrusion detection

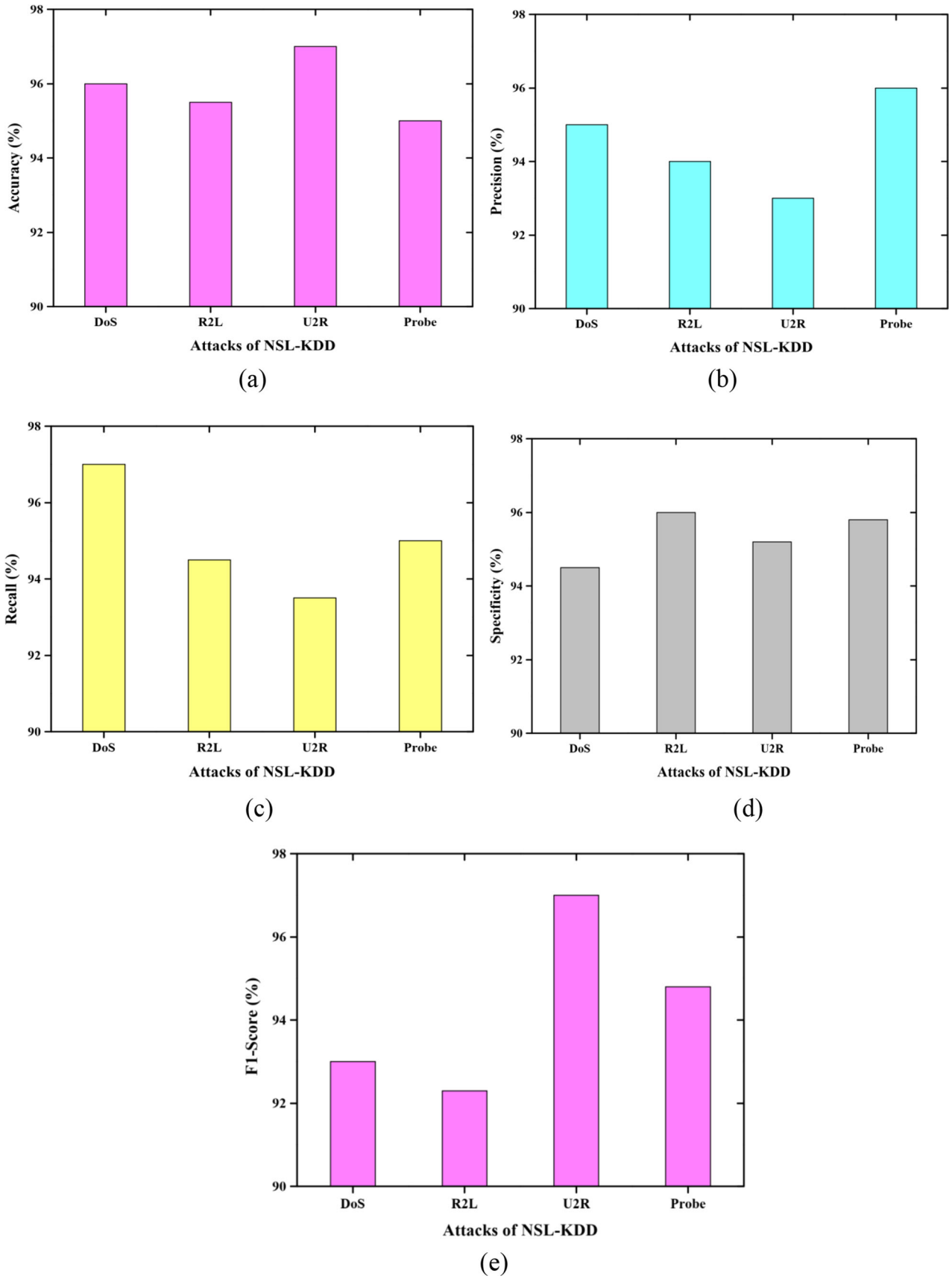

Fig. 4 Performance analysis of various attacks in NSL-KDD dataset. **a** Accuracy, **b** precision, **c** recall, **d** specificity, **e** F1-score

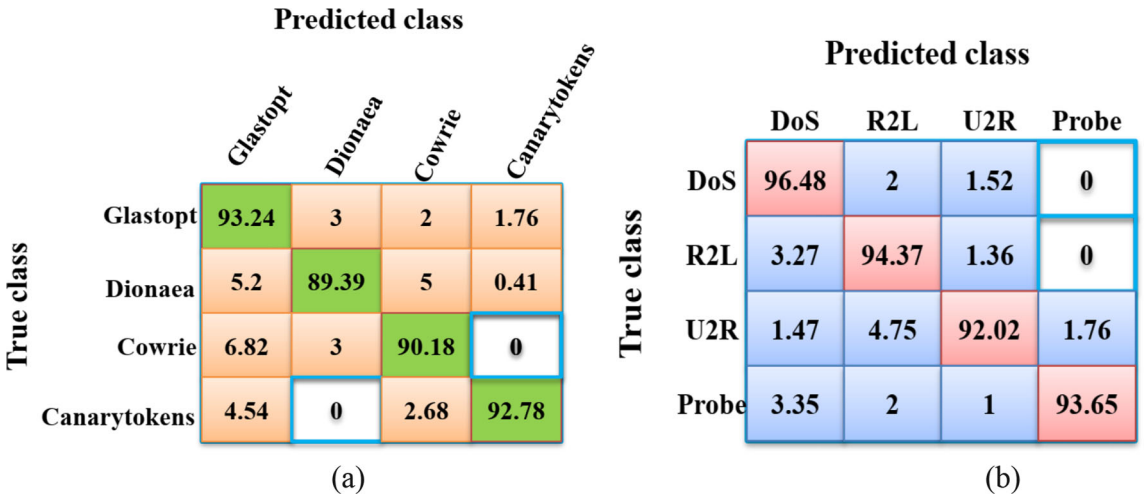

Fig. 5 Confusion matrix, **a** honeypot dataset, **b** NSL-KDD dataset

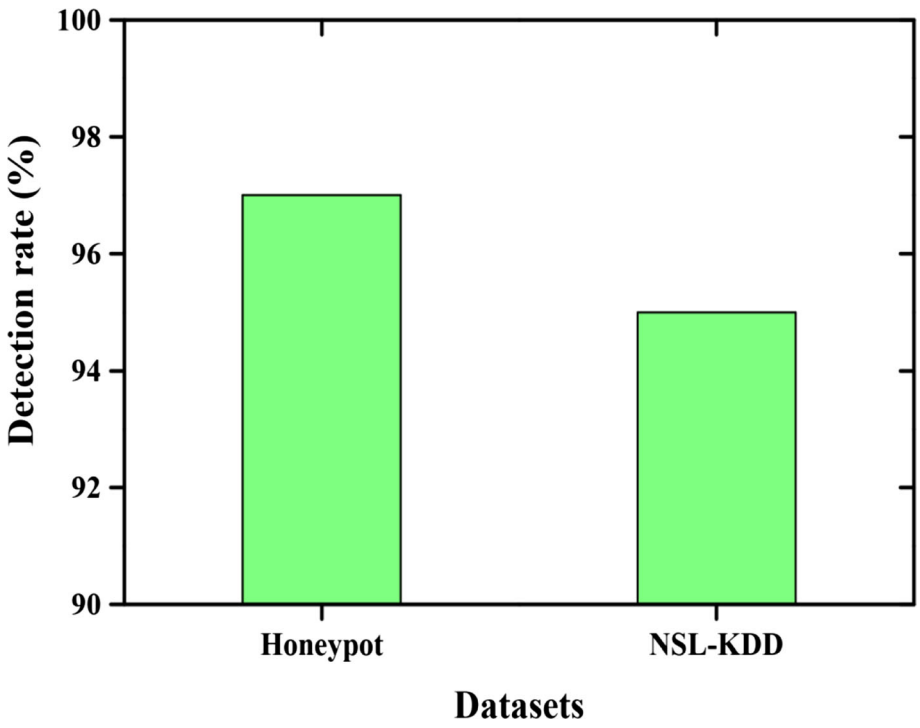

Fig. 6 Detection rate analysis

model uses the krill herd optimization (KHO) algorithm to effectively select or extract the significant features from the databases and then classified using BiLSTM based explainable AI (BiLSTM-XAI) framework. This framework offers effective monitoring of industry 4.0 applications against intrusions such as abnormal behaviors and unauthorized access. The implementation of XAI models reduces the complexities of BiLSTM framework and thus enhances the detection accuracy by providing explanations for each learning prediction. The proposed method was examined by two datasets, namely Honeypot and NSL-KDD, respectively. The performance of the proposed BiLSTM-XAI method is higher and the NSL-KDD dataset achieves better performance than dataset. The performance metrics such as accuracy (A), precision (P), specificity (SP), F1-score and Recall (R) are applied to evaluate the efficiency of the proposed method. The classification accuracy of the proposed BiLSTM-XAI method is higher compared to federated deep learning mechanism (DeepFed), artificial intelligence enabled multimodal fusion-based intrusion detection system (AIMMF-IDS), experience driven attack design with federated learning

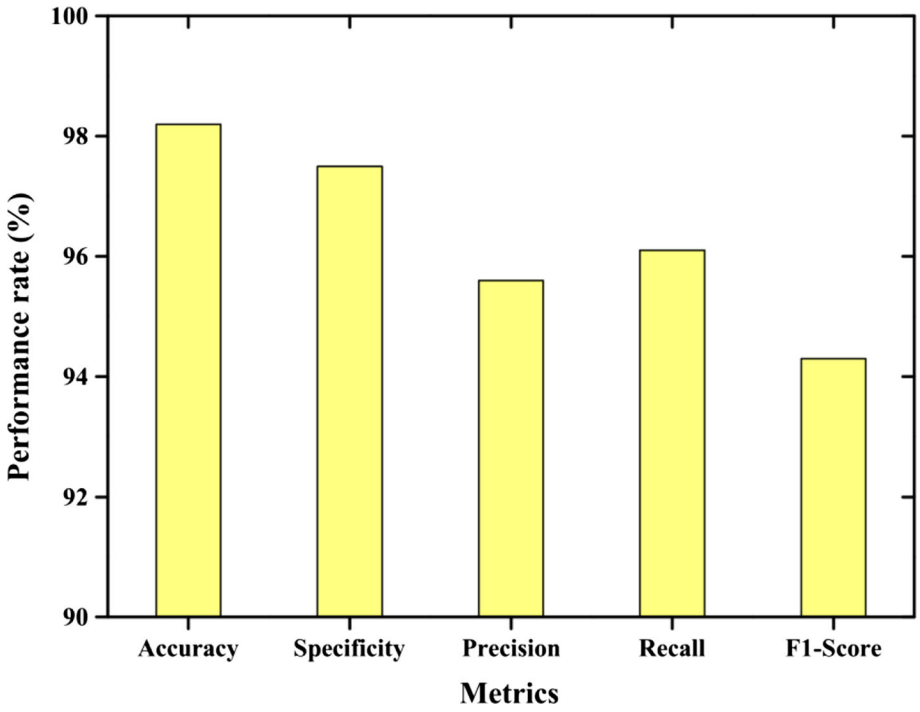

Fig. 7 Performance metrics analysis of the proposed BiLSTM-XAI method

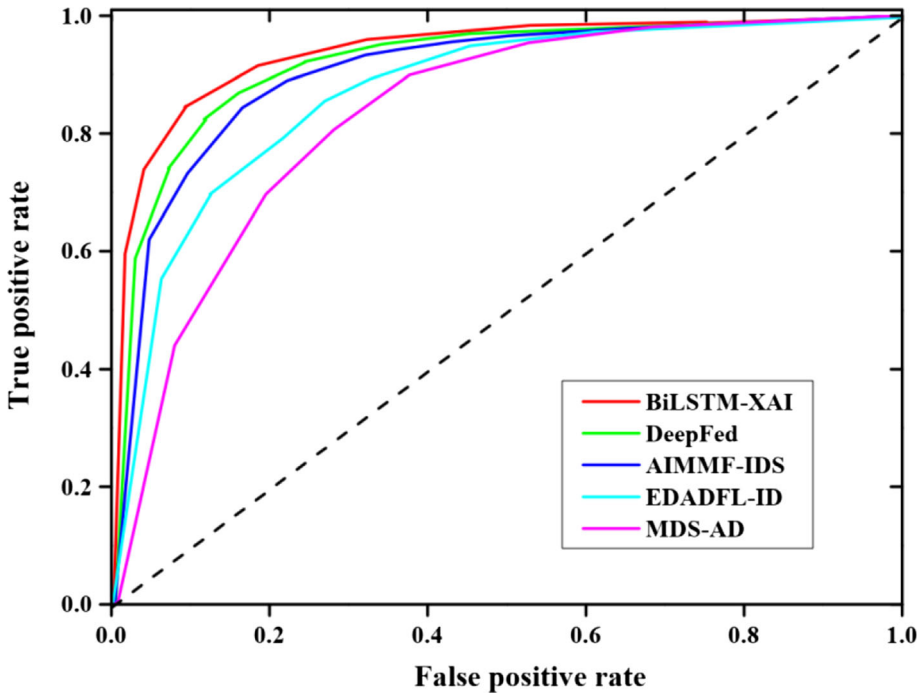

Fig. 8 ROC curve analysis

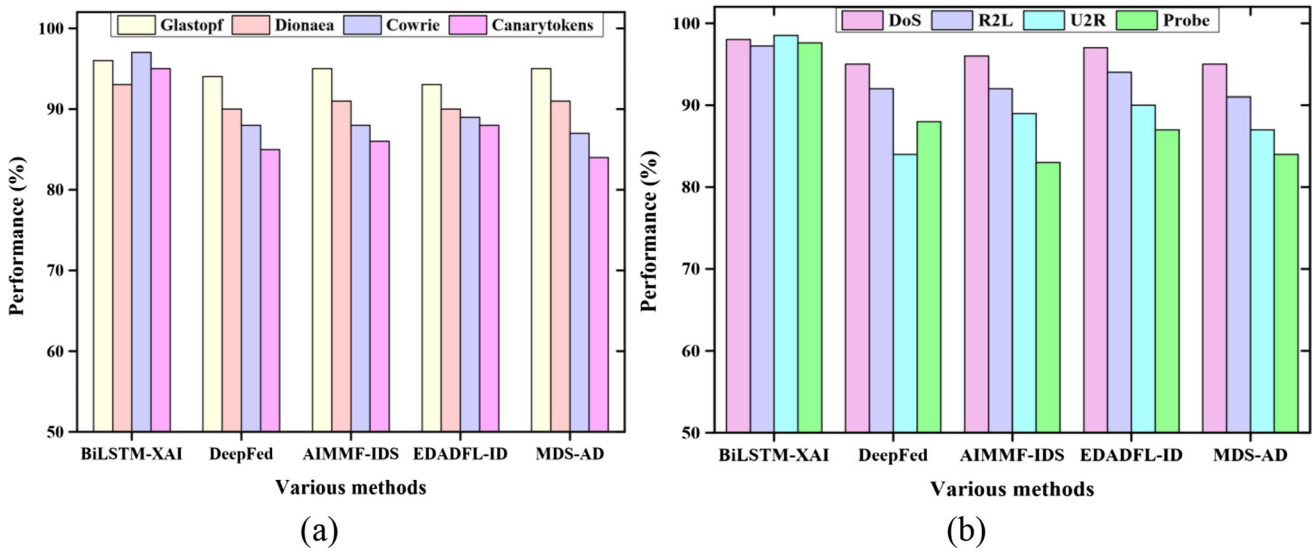

Fig. 9 Performance analysis of **a** honeypot dataset, **b** NSL-KDD dataset

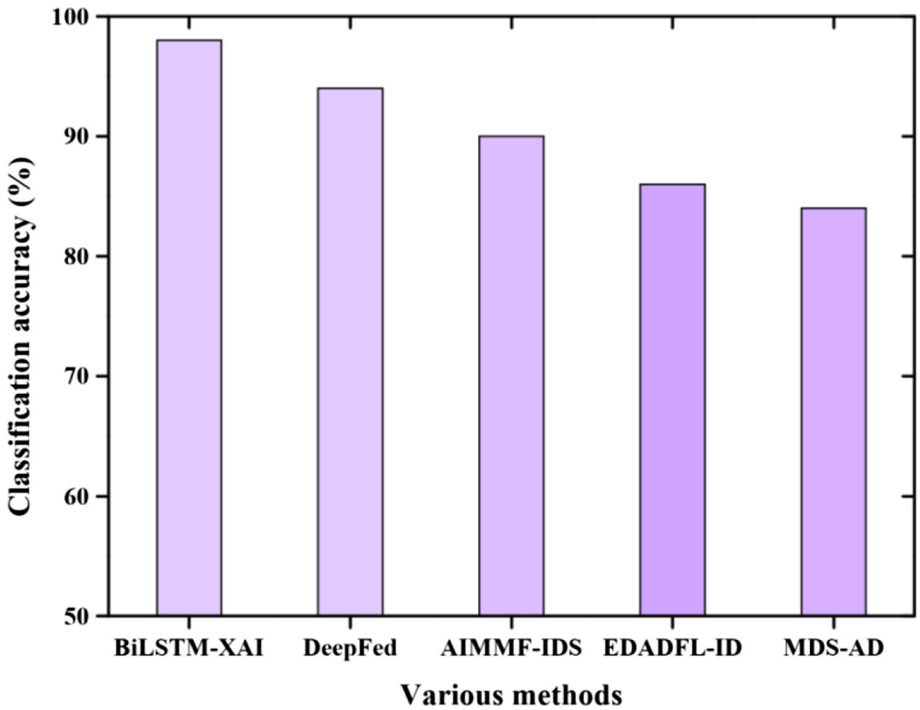

Fig. 10 Classification accuracy analysis

based intrusion detection (EDADFL-ID) and Multi-Aspect Data Streams-anomaly detection (MDS-AD). The obtained detection rate of Honeypot is 97.2% and the NSL-KDD dataset is 95.8%, respectively. In future, the unknown adversarial attacks in the network will be determined effectively using some other hybrid metaheuristic approaches.

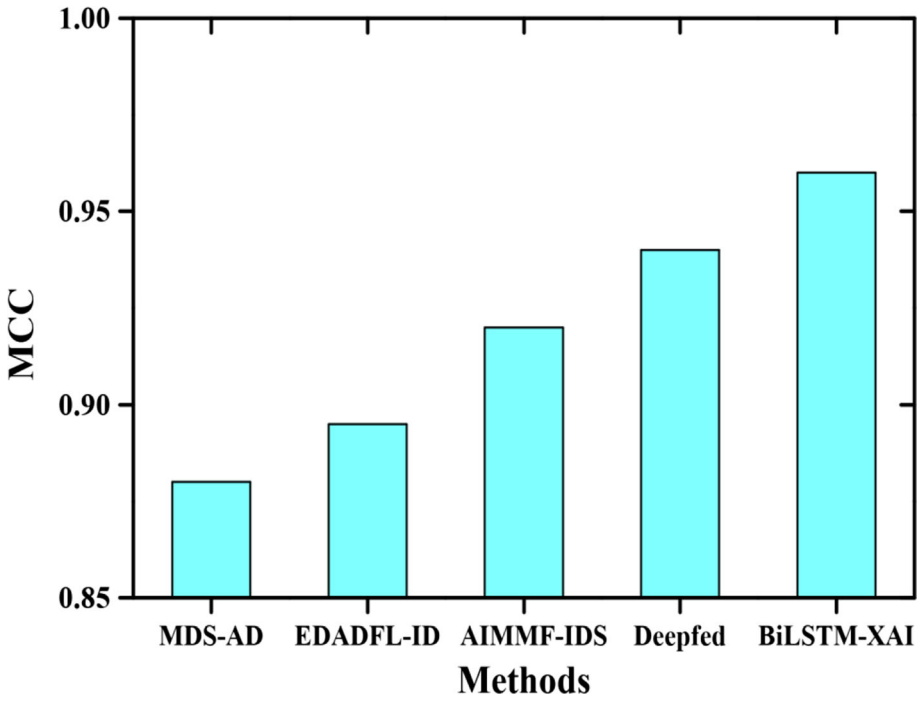

Fig. 11 MCC analysis

Author contributions SS agreed on the content of the study. SS and SSS collected all the data for analysis. SS agreed on the methodology. SS and SSS completed the analysis based on agreed steps. Results and conclusions are discussed and written together. Both author read and approved the final manuscript.

Funding Not applicable.

Data availability Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Code availability Not applicable.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Consent to participate Not applicable.

Consent for publication Not applicable.

Human and animal rights This article does not contain any studies with human or animal subjects performed by any of the authors.

Informed consent For this type of study, informed consent is not required.

References

1. Bajic B, Rikalovic A, Suzic N, Piuri V (2020) Industry 4.0 implementation challenges and opportunities: a managerial perspective. IEEE Syst J 15(1):546–559
2. Wanasinghe TR, Trinh T, Nguyen T, Gosine RG, James LA, Warrian PPJ (2021) Human centric digital transformation and operator 4.0 for the oil and gas industry. IEEE Access 9:113270–113291
3. Ferrag MA, Shu L, Djallel H, Choo KKR (2021) Deep learning-based intrusion detection for distributed denial of service attack in agriculture 4.0. Electronics 10(11):1257
4. Zheng T, Ardolino M, Bacchetti A, Perona M (2021) The applications of Industry 4.0 technologies in manufacturing context: a systematic literature review. Int J Prod Res 59(6):1922–1954
5. Kiran MB (2021) Significance of intruder detection techniques in the context of industry 4.0. In: Proceedings of the international conference on industrial engineering and operations management. pp 2977–2985
6. Gunduz MZ, Das R (2020) Cyber-security on smart grid: Threats and potential solutions. Comput Netw 169:107094
7. Ahmad I, Shah SAA, Al-Khasawneh MA (2021) Performance Analysis of Intrusion Detection systems for smartphone security enhancements. In: 2021 2nd international conference on smart computing and electronic enterprise (ICSCEE), pp 19–25, IEEE
8. Sun M, Li X, Yang R, Zhang Y, Zhang L, Song Z, Liu Q, Zhao D (2020) Comprehensive partitions and different strategies based on ecological security and economic development in Guizhou Province. China J Clean Prod 274:122794
9. Saghezchi FB, Mantas G, Violas MA, de Oliveira Duarte AM, Rodriguez J (2022) Machine learning for DDoS attack detection in industry 4.0 CPPSs. Electronics 11(4):602
10. Saxena N, Hayes E, Bertino E, Ojo PP, Choo KKR, Burnap PP (2020) Impact and key challenges of insider threats on organizations and critical businesses. Electronics 9(9):1460
11. Alohali MA, Al-Wesabi FN, Hilal AM, Goel S, Gupta D, Khanna A (2022) Artificial intelligence enabled intrusion detection systems for cognitive cyber-physical systems in industry 4.0 environment. Cognit Neurodyn 16:1–13
12. Tahir B, Jolfaei A, Tariq M (2021) Experience driven attack design and federated learning based intrusion detection in industry 4.0. IEEE Trans Ind Inf 18:6398–6405
13. Qi L, Yang Y, Zhou X, Rafique W, Ma J (2021) Fast anomaly identification based on multi-aspect data streams for intelligent

intrusion detection toward secure industry 4.0. IEEE Trans Ind Inf 18:6503–6511
14. Yang K, Shi Y, Yu Z, Yang Q, Sangaiah AK, Zeng H (2022) Stacked one-class broad learning system for intrusion detection in industry 4.0. IEEE Trans Ind Inf 19:251–260
15. Ibitoye O, Shafiq O, Matrawy A (2019) Analyzing adversarial attacks against deep learning for intrusion detection in IoT networks. In: 2019 IEEE global communications conference (GLOBECOM), pp. 1–6. IEEE
16. Saghezchi FB, Mantas G, Violas MA, de Oliveira Duarte AM, Rodriguez J (2022) Machine learning for DDoS attack detection in industry 4.0 CPPSs. Electronics 11(4):602
17. Khoa TV, Saputra YM, Hoang DT, Trung NL, Nguyen D, Ha NV, Dutkiewicz E (2020) Collaborative learning model for cyberattack detection systems in iot industry 4.0. In: 2020 IEEE wireless communications and networking conference WCNC, pp. 1–6. IEEE.
18. Li B, Wu Y, Song J, Lu R, Li T, Zhao L (2020) DeepFed: Federated deep learning for intrusion detection in industrial cyber–physical systems. IEEE Trans Industr Inf 17(8):5615–5624
19. Chowdhury D, Poddar S, Banarjee S, Pal R, Gani A, Ellis C, Arya RC, Gill SS, Uhlig S (2022) CovidXAI: explainable ai-assisted web application for COVID-19 vaccine prioritisation. Int Technol Lett. https://doi.org/10.1002/itl2.381pp.e381
20. Krishnaveni S, Sivamohan S, Sridhar S, Prabhakaran S (2022) Network intrusion detection based on ensemble classification and feature selection method for cloud computing. Concurr Comput Pract Exp 34(11):e6838
21. Krishnaveni S, Sivamohan S, Sridhar SS, Prabakaran S (2021) Efficient feature selection and classification through ensemble method for network intrusion detection on cloud computing. Clust Comput 24(3):1761–1779
22. Barnard PP, Marchetti N, DaSilva LA (2022) Robust network intrusion detection through explainable artificial intelligence (XAI). IEEE Netw Lett 4(3):167–171
23. Liu H, Zhong C, Alnusair A, Islam SR (2021) FAIXID: a framework for enhancing ai explainability of intrusion detection results using data cleaning techniques. J Netw Syst Manage 29(4):1–30
24. Larriva-Novo X, Villagrá VA, Vega-Barbas M, Rivera D, Sanz Rodrigo M (2021) An IoT-focused intrusion detection system approach based on preprocessing characterization for cybersecurity datasets. Sensors 21(2):656
25. Li X, Yi PP, Wei W, Jiang Y, Tian L (2021) LNNLS-KH: a feature selection method for network intrusion detection. Sec Commun Netw. https://doi.org/10.1155/2021/8830431
26. Resma KB, Nair MS (2021) Multilevel thresholding for image segmentation using Krill Herd optimization algorithm. J King Saud Univ-Comput Inf Sci 33(5):528–541
27. Abdul W, Alsulaiman M, Amin SU, Faisal M, Muhammad G, Albogamy FR, Bencherif MA, Ghaleb H (2021) Intelligent real-time Arabic sign language classification using attention-based inception and BiLSTM. Comput Electr Eng 95:107395
28. Knapič S, Malhi A, Saluja R, Främling K (2021) Explainable artificial intelligence for human decision support system in the medical domain. Mach Learn Knowl Extractio 3(3):740–770
29. Kwon D, Natarajan K, Suh SC, Kim H, Kim J (2018) An empirical study on network anomaly detection using convolutional neural networks. In: ICDCS pp 1595–1598