applied sciences

MDPI

Article

Explainable Artificial Intelligence Enabled Intrusion Detection Technique for Secure Cyber-Physical Systems

Latifah Almuqren [1], Mashael S. Maashi [2], Mohammad Alamgeer [3], Heba Mohsen [4], Manar Ahmed Hamza [5,*] and Amgad Atta Abdelmageed [5]

[1] Department of Information Systems, College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia
[2] Department of Software Engineering, College of Computer and Information Sciences, King Saud University, P.O. Box 103786, Riyadh 11543, Saudi Arabia
[3] Department of Information Systems, College of Science & Art at Mahayil, King Khalid University, Abha 62529, Saudi Arabia
[4] Department of Computer Science, Faculty of Computers and Information Technology, Future University in Egypt, New Cairo 11835, Egypt
[5] Department of Computer and Self Development, Preparatory Year Deanship, Prince Sattam bin Abdulaziz University, Al-Kharj 16278, Saudi Arabia
* Correspondence: ma.hamza@psau.edu.sa

Abstract: A cyber-physical system (CPS) can be referred to as a network of cyber and physical components that communicate with each other in a feedback manner. A CPS is essential for daily activities and approves critical infrastructure as it provides the base for innovative smart devices. The recent advances in the field of explainable artificial intelligence have contributed to the development of robust intrusion detection modes for CPS environments. This study develops an Explainable Artificial Intelligence Enabled Intrusion Detection Technique for Secure Cyber-Physical Systems (XAIID-SCPS). The proposed XAIID-SCPS technique mainly concentrates on the detection and classification of intrusions in the CPS platform. In the XAIID-SCPS technique, a Hybrid Enhanced Glowworm Swarm Optimization (HEGSO) algorithm is applied for feature selection purposes. For intrusion detection, the Improved Elman Neural Network (IENN) model was utilized with an Enhanced Fruitfly Optimization (EFFO) algorithm for parameter optimization. Moreover, the XAIID-SCPS technique integrates the XAI approach LIME for better understanding and explainability of the black-box method for accurate classification of intrusions. The simulation values demonstrate the promising performance of the XAIID-SCPS technique over other approaches with maximum accuracy of 98.87%.

Keywords: security; intrusion detection; cyber-physical systems; explainable artificial intelligence; feature selection

Citation: Almuqren, L.; Maashi, M.S.; Alamgeer, M.; Mohsen, H.; Hamza, M.A.; Abdelmageed, A.A. Explainable Artificial Intelligence Enabled Intrusion Detection Technique for Secure Cyber-Physical Systems. *Appl. Sci.* **2023**, *13*, 3081. https://doi.org/10.3390/app13053081

Academic Editor: Luis Javier Garcia Villalba

Received: 16 January 2023
Revised: 12 February 2023
Accepted: 17 February 2023
Published: 27 February 2023

1. Introduction

A cyber-physical system (CPS) enables physical gadgets with sensing abilities to interact with the internet or controllers as required [1]. The communication channels used can be short-distance communication or wireless technology to continually upgrade the physical environment condition or physical device status to a remote server or controller [2]. The recent developments in wireless communications and sensor technologies have utilized the CPS in several application fields, such as aviation and chemical industries, electronics, material manufacturing with automatic supply chain, and smart industries, which includes transport, etc. [3]. The advent of CPS applications in different fields even paves the way for novel security challenges and issues to safeguard confidential data or infrastructure against cybersecurity. The attacks include cyberattacks via internet-connected devices, and physical assaults which can lead to supply chain disruption or system failures [4]. Therefore, CPS

security is very challenging compared to the classical IT and network substructure [5]. Figure 1 represents the CPS and potential security threats.

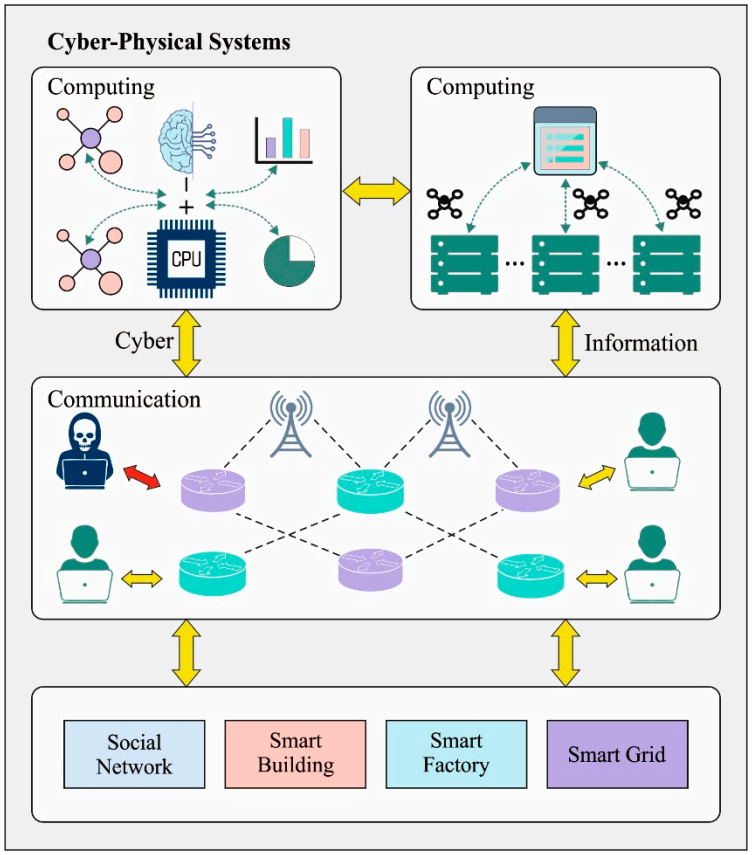

Figure 1. CPS and potential security threats.

Intrusion detection systems (IDS) are highly capable of detecting and preventing data breaches by stopping and detecting intrusions [6–8]. Anomaly detection (AD) and Misuse detection are two kinds of ID. Misuse detection depends on patterns or information, whereas AD relies on behavior. Present IDS have a higher detection rate, which leads to several false alarms. In an IDS, false positives must be reduced. Several IDS were applied with the help of various ML methods as they can find valuable data from the database. Such techniques have the potential to diminish false positives. ML techniques, involving IDS, an association of rules, and GA were enforced through artificial neural networks (ANNs). Ensemble learning merges different ML techniques. The authors found that an ensemble method involving ML techniques lessens false positives.

This study develops an Explainable Artificial Intelligence Enabled Intrusion Detection Technique for Secure Cyber-Physical Systems (XAIID-SCPS). In the XAIID-SCPS technique, a Hybrid Enhanced Glowworm Swarm Optimization (HEGSO) algorithm is applied for feature selection purposes. For intrusion detection, the Improved Elman Neural Network (IENN) method was exploited with an Enhanced Fruitfly Optimization (EFFO) algorithm for parameter optimization. Moreover, the XAIID-SCPS technique integrates the XAI approach LIME for better explainability and understanding of the black-box method for the accurate classification of intrusions. The simulation values of the XAIID-SCPS technique can be tested on benchmark intrusion datasets.

The rest of the study Is organized as follows. Section 2 elaborates on the existing IDS models in the CPS environment. Next, the proposed XAIID-SCPS technique is discussed in Section 3, and the experimental results are defined in Section 4. Finally, Section 5 concludes the work with key findings and possible future work.

2. Related Works

Radanliev et al. [9] presented a new mathematical approach for integrating concepts for cognition engine design, edge computing and Artificial Intelligence and Machine Learning to automate anomaly detection. The authors in [10] identified a wide range of methods for cyber analytics and explore the risks of deliberately influencing or disrupting behaviors in socio-technical systems. It modeled the connections and interdependencies between a system's edge components to both external and internal services and systems. Munir et al. [11] presented an AI-based exploratory cyber-physical safety analyzer structure. This structure modelled supervised learning-related AI techniques such as K-KNN, DT, LR, LSTM, and RF to detect and predict spoofing attacks and cyber jamming. Afterwards, the developed structure examines the conditional dependency depending on Pearson's correlation coefficient between controlled messages to identify the reason for effective assaults related to the outcome of the AI system. Colelli et al. [12] intended to offer a tool for detecting cyberattacks in CPS. This technique depends on ML for enhancing the system's security. It is possible to assess the classifier performance of the three methods through analysis of values assumed by ML. The method gained with the help of trained set permits categorizing a sample of anomalous behavior and a sample that can be relevant to usual behavior. In [13], a novel risk assessment technique was presented in this study for quantifying the effect of malicious assaults on the physical mechanism of ICPS. This technique helps to perform suitable attack mitigation measures. The technique leverages a Bayesian network to devise an attack propagation procedure and infer the probability of actuators and sensors being compromised.

Schneider and Böttinger [14] introduced a uniform framework and method to apply an AD to various Fieldbus protocols. The author used stacked denoising AE to derive a packet classification and feature learning technique in one step. In addition, the author created a potential structure which can even manage the increased amount of transmission in CPS. Sharma et al. [15] devised superior lightweight behavior rule specification-related misbehavior recognition for IoT-embedded CPS (BRIoT). The main aim of this technique was to devise a mechanism with which misbehavior of an IoT gadget was established so assaults exploiting the exposed susceptibility can be identified with the use of formal verification and automated model checking, irrespective of whether the assault was unknown or known.

In [16], a security decision-making technique related to the stochastic game model was modelled for characterizing the communication among defenders and attackers in ICPS, producing the best defense approaches for reducing system losses. The major difference in this technique was that it offers a practical means to frame a cross-layer security game method for ICPS with time-based unified payoff quantification and quantitative vulnerability analysis. Huang and Zhu [17] presented a dynamic game structure to devise long-term communication between a proactive defender and a stealthy attacker. The deceptive and stealthy behaviors were captured by a multi-stage game of unfinished data, where all players have their private data unknown to others.

Wang et al. [18] presented a Knowledge Distillation method dependent upon Triplet CNN for improving the model efficiency and significantly improve the speed of AD for industrial CPS (ICPS) and decrease the complexity of model. A novel NN trained approach termed as K-fold cross training was also presented for enhancing the accuracy of AD. Tang et al. [19] examine an ICPS with IDS dependent upon the Diffusion model. Primarily, data equivalent to the rare class can be created by the diffusion model that generates the trained database of various classes balanced.

Ramadevi et al. [20] introduced an Optimal DBN-based distributed IDS (ODBN-IDS) for securing CPS platform. The presented method was focused on pre-processing the cloud network traffic data and improve their quality to the next level. An equilibrium Optimizer Algorithm (EOA) was utilized for fine-tuning the hyperparameters in DBN technique. Alohali et al. [21] examined a novel AI-assisted multi-modal fusion-based IDS (AIMMF-IDS) for CCPS in industry 4.0 platform. Moreover, an improved fish swarm optimizer

based FS (IFSO-FS) approach was utilized for suitable FSs. Dutta et al. [22] presented a robust AD method dependent upon semi-supervised ML approach permitting us near real-time localization of attacks. A DNN structure was utilized for AD dependent upon reconstruction error.

3. The Proposed Model

In this study, we have presented an automated intrusion detection technique, named the XAIID-SCPS technique, for the CPS environment. The presented XAIID-SCPS method mostly concentrates on the detection and classification of intrusions in the CPS platform. In the XAIID-SCPS technique, several subprocesses are involved, namely, data pre-processing, HEGSO-based feature selection, IENN-based classification, and EFFO-based parameter tuning. Figure 2 illustrates the overall flow of the XAIID-SCPS system.

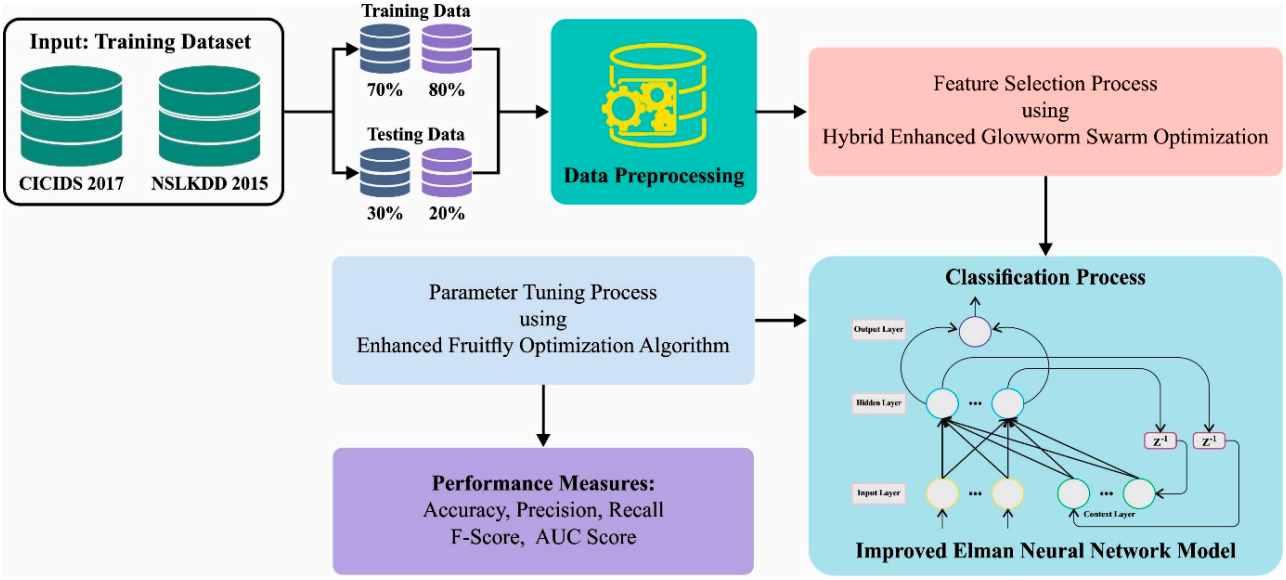

Figure 2. Overall flow of XAIID-SCPS system.

3.1. Preprocessing

In this phase, data conversion and data normalization are the two stages of preprocessing. The input dataset in .xls form is converted into .csv form during data conversion. In addition, data normalization can be performed by the min–max technique, where the minimum and maximum values are taken from the given dataset. It aims to the normalization of sample to a higher value of 1 and a lower value of 0 as follows:

$$Min - Max.Norm = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

3.2. Feature Selection Using HEGSO Algorithm

To select features accurately, the HEGSO algorithm is exploited in this study. The classical GSO technique is based on glowworm behaviors and was rooted in the natural activity of glowworms during the night [23]. The glowworm exhibits a kind of communication with other glowworms in a group depending on the luciferin. If the luciferin was high, then the light produced by the glowworm was high. As a result, the glowworm goes towards it. In this neighborhood, Luciferin update, movement, and update phases are the three different stages of the GSO technique.

The luciferin update stage refers to luciferin creation. The quantity of luciferin was directly proportional to the fitness of the current site on the main function. During the movement stage, the glowworm selects the use of probabilistic means to go towards the neighbor where the value is high. The succeeding stage involves the adaptive neighborhood

range for perceiving the peak presence. The optimizing process can be performed by using the HEGSO technique. The presented study applies the hybrid model where the HEGSO technique is used. Moreover, the usage of two genetic algorithms includes mutation and crossover operators. The optimization algorithm is initiated when the path is found between the beginning and the end points of the vehicular nodes that are presented in the road segment. The HEGSO method can be given in the following,

Step 1: Initialize the 'Eearch Agent' (EA), namely, separate glowworm X and the population size can be represented as Q. z shows the step size, the maximum number of iterations is represented by N_{it}, I_o indicates the initial value of luciferin, r_o indicates the primary value of radial gamut of SA, and n denotes the hour's instance.

$$X = \{x_1,\ x_2,\ \ldots .\ x_Q\} \tag{2}$$

Step 2: Calculate the fitness function using the following expression:

$$F = \sum_{r=1}^{n} \frac{L}{EV_t(V_i)} \tag{3}$$

where L denotes the overall length of the road segment, the average velocity of the vehicle on road was represented as EV_t and V_i specify the vehicle.

Step 3: In the luciferin stage, every new SA is calculated by the following equation:

$$l_j(n+1) = (1-\rho)l_j(n) + \beta J_j(n+1) \tag{4}$$

where $l(n)$ indicates the luciferin value of SA at n time, $l_j(n+1)$ indicates the luciferin value of SA at $(n+1)$ time, J_j denotes fitness function, and β denotes luciferin decay constant with the value in gamut [0, 1].

Step 4: The main function of every novel SA is evaluated by a similar fitness function that was declared in step 2.

Step 5: Now, the SA headed for the neighboring glowworm was based on luminance. The SA movement can be determined by Euclidean distance among the glowworms. Here, SA heads for evaluated neighborhoods based on the probabilistic model, which can be expressed as follows.

$$x_j(n+1) = x_j(n) + Z\left(\frac{xk(n) - x_j(n)}{\|xk(n) - x_j(n)\|}\right) \tag{5}$$

In Equation (5), Z denotes step size.

Step 6: Here, the decision range was upgraded, namely, the neighborhood range of SA is upgraded as follows:

$$r_d(n+1) = \min\left(r_s,\ \max\left(O,\ r_d(n)\right) + B\left(n_e - \left|N_j(n)\right|\right)\right) \tag{6}$$

In Equation (6), B indicates the constant, r_s represent the largest sensing radius of the glowworm. n_e shows the SA having a higher luciferin value in the decision range, $r_d(n+1)$ and rd denote the upgraded and preceding values of the neighborhood range.

Step 7: Check whether the ending condition is met. When the ending condition is satisfied, then a better solution can be attained. When the ending conditions are not satisfied after two genetic operators, namely, crossover and mutation operators, are used. Furthermore, the 2-point crossover was used. This can be followed by a mutation process where there are specific genes. The value that can be chosen for the crossover operator were mutated and then integrated.

In the proposed HEGSO approach, the fitness function is intended to have a balance between the classification accuracy (maximum) and the number of selected features in

all the solutions (minimum) attained Equation (10) characterizes the fitness function to estimate the solution.

$$Fitness = \alpha\gamma_R(D) + \beta\frac{|R|}{|C|} \tag{7}$$

where $\gamma_R(D)$ signifies the classifier error rate of the given classifier. $|R|$ indicates the cardinality of the selective subset and $|C|$ denotes the overall number of features in data, and α and β denote the two parameters corresponding to the significance of classifier quality and subset length, respectively. $\in [1, 0]$ and $\beta = 1 - \alpha$ as explained in Algorithm 1.

Algorithm 1. HEGSO Algorithm

Process HEGSO
Start: Input: X_i, Q, z, N_{it}, I_o, r_o.
Define fitness function
While $t < N_{iy}$ do
For every glowworm do

$$l_j(n+1) = (1-\rho)l_j(n) + \beta J_j(n+1)$$
$$x_j(n+1) = x_j(n) + Z\left(\frac{xk(n)-x_j(n)}{\|xk(n)-x_j(n)\|}\right)$$
$$r_d(n+1) = \min\left(r_s, \max\left(O, r_d(n) + B(n_e - N_j(n)\right)\right)$$

End for
End while
Implement the crossover and mutation
Return X_{best}
End Process

3.3. Intrusion Detection Using Optimal IENN Model

For detecting intrusions accurately, the IENN model is used. The ENN was a fully connected dynamic feedback NN that has a local feedback function and local memory unit. It realizes the mapping of a dynamic system and realizes the modeling of a static system and directly reflects the dynamic features of the model [24]. In comparison to FFNN, Elman was a one-step delay operator that attains the purpose of short-term storage, and add another receiving layer based on a three-layer structure of hidden, input and resultant layers. Thus, the Elman method can adapt to time-varying features. Simultaneously, it has strong network stability and computing power. While there is a non-linear relationship between the parameters of SOH and LIB, such features of the ENN enable -linear relation with better precision. Thus, the ENN is selected as an infrastructure.

The ENN architecture with one output unit, two input units and three hidden units. k denotes the k^{th} time; $X(k)$ signifies the input vector of an input layer; w_{ij}^1, w_{ij}^2, and w_{ij}^3 signify the weight connection from input to hidden layers (HLs) from receiving to HL and from HL to the output layer, correspondingly. $C(k)$, $C^*(k)$, and $Y(k)$ signify the output vector of the hidden, getting, and resultant layers, correspondingly. $b1$ to $b4$ denotes the threshold of getting and resultant layers, correspondingly, and it can be mathematically expressed as follows:

$$C^*(k) = C(k-1) \tag{8}$$

$$C(k) = f\left(w_{ij}^1 X(k) + bi + w_{ij}^2 C^*(k)\right) \tag{9}$$

$$Y(k) = g\left(w_{ij}^3 C(k) + bi\right) f(\cdot) \tag{10}$$

Here, bi indicates the corresponding threshold, $g(\cdot)$ denotes the activation function of output neurons that is usually a linear integration, and f signifies the activation function of HL.

However, ENNs have their benefits and drawbacks. Focusing on the drawbacks of NN algorithms in real-time engineering applications, research workers have primarily investigated and amended from two factors of learning algorithms and network topology. The amendment of NN topology will dramatically increase the computation problem

that was not advantageous to the development of NN performance. Thus, the additional momentum technique improves the learning method of the NN and the algorithm implies that during backpropagation, a part of the preceding weight change was included to present weight adjustment values by using momentum factors and utilized as an actual weight adjustment values:

$$\Delta w(t+1) = \alpha \Delta w(t) + (1-\alpha)\eta \frac{\partial E(t)}{\partial w(t)} \tag{11}$$

$$\Delta w(t+1) = w(t) + \Delta w(t+1) \tag{12}$$

Now, t denotes the count of training times, α indicates the momentum factor, usually fixed to 0.95, and η indicates the learning rate.

To alter the variables related to the IENN method, the EFFO approach was used. In general, the FFO algorithm is based on the behaviors of fruit flies during the food search process [25]. This method comprises three stages:

(1) Initializing stage: the fruit flies have dispersed arbitrarily as X_axis and Y_axis, where rv denotes the uniform distribution random number.

$$X_i = X_axis + rv \tag{13}$$

$$Y_i = Y_axis + rv \tag{14}$$

(2) Path Construction stage: the distance and odor intensity of each fruit fly is performed by using the following equation.

$$Distance_i = \sqrt{X_i^2 + Y_i^2} \tag{15}$$

$$SM_i^c = \frac{l}{Distance_i} \tag{16}$$

where $Distance_i$ denotes the distance between the food location and the i^{th} individual, SM_i^c denotes the judgment value of odor intensity concentration, and this was the reciprocal of distance.

(3) Fitness evaluation stage: The fitness formula can be expressed in the following.

$$smell_i = function\ (SM_i^c) \tag{17}$$

$$smell^{best}, index^{best} = \max\ (smell_i) \tag{18}$$

where $smell_i$ shows the value of odor intensity of the distinct fruit flies, $smell^{best}$ and $index^{best}$ denote the maximum component and its corresponding indices have dissimilar dimensions of smell vector, and max $(smell_i)$ indicates the highest odor intensity concentration amongst fruit flies.

(4) Movement phase: The fruit fly provides a better value of odor intensity and flies towards that position correspondingly. The pseudocode of the FFO algorithm (Algorithm 2) is shown below.

$$BEST\ SMELL = smell^{best} \tag{19}$$

$$X_axis = X\left(index^{best}\right) \tag{20}$$

$$Y_axis = Y\left(index^{best}\right) \tag{21}$$

The EFFO algorithm is defined by the incorporation of a chaotic concept. The chaotic signal made by the deterministic system has the quality of genus-randomness. The curve can be defined by the initial value and chaos mapping parameter.

Algorithm 2. FFOA algorithm

Step 1 Initializes the parameter
Step 2 Repeat
Select the random position under distance and odor intensity
Assess the fitness function SM_i^c
Recognize the fruit fly with the highest odor intensity concentration between
The fruit fly swarm
Ranking of solutions, and upgrading the better solution
Step 3 Return the better solution

Logistic mapping was leveraged widely. The Logistic chaotic mechanism has complicated dynamical behavior and it is defined as follows:

$$\lambda_{i+1} = \mu \times \lambda_i \times (1 - \lambda_i) \tag{22}$$

$\lambda \in [0,\ 1], i = 0, 1, 2, \cdots, \mu$ is in [1, 4]. The study recommended that μ was closer to 4, and λ was closer to the average distribution between [0, 1]. Meanwhile, the system was completely chaotic if μ value is 4. The initial population was a significant part of the intelligent optimizer technique that affect the final solution quality and convergence rate. Logistic chaotic mapping was leveraged for initializing the population of FFO that exploits the solution space to enhance the efficiency of the model.

The fitness selection is a critical factor in the EFFO algorithm. Solution encoding can be used to assess the aptitude (goodness) of the candidate solution. Here, the accuracy value is the main condition used to design a fitness function.

$$Fitness = \max(P) \tag{23}$$

$$P = \frac{TP}{TP + FP} \tag{24}$$

From the expression, TP represents the true positive and FP denotes the false positive value.

3.4. Modeling of XAI Using LIME

In this work, the XAIID-SCPS technique integrates the XAI approach LIME for superior explainability and understanding of the black-box method for the accurate classification of intrusions [26]. Local interpretable model-agnostic explanation (LIME) could describe several ML techniques for regression prediction, utilizing the featured value change in the data samples to convert the featured value into the contribution of the predictor. The explainer provides a local interpretation of the data samples. For instance, interpretable models in LIME frequently use decision trees (DTs) or linear regression (LR) that are trained by the small perturbation (hiding parts of an image, adding random noise, and removing specific words) in the model. The quality of this model seems to be increasing and was utilized to solve the best part of business victimization data. Moreover, there was a persistent tradeoff between interpretability and model accuracy. In general, the accuracy is enhanced to use sophisticated techniques such as random forest, material, boosting, SVM, and call trees, which are "blackbox" techniques. The local interpretable model agnostic explanation (LIME) gives a clear description of the problem with the blackbox classifier. The LIME is a way to understand an ML BlackBox technique by perturbing the input dataset and seeing how the prediction changes. The LIME is used for any ML black-box models. The main steps are given below:

In the class explain_instance, a technique named explain_instance accepts the reference to the instance where the explanation is required, together with the number of features to be added in the explanation and the trained model's prediction technique.

A TabularExplainer can be initialized by the data used for training the data about the features, and various class names.

4. Performance Validation

In this section, the intrusion detection results of the XAIID-SCPS method can be investigated on two datasets namely NSLKDD2015 and CICIDS2017 dataset as shown in Table 1.

Table 1. Details dataset.

Class	No. of Samples	
	NSLKDD 2015	CICIDS 2017
Normal	67,343	50,000
Anomaly	58,630	50,000
Total Number of Samples	**125,973**	**100,000**

The confusion matrices of the XAIID-SCPS method on the NSLKDD 2015 dataset are illustrated in Figure 3. The results recognized the normal and abnormal samples proficiently. For instance, with 80% of TRS, the XAIID-SCPS technique identifies 53,611 normal and 46,054 abnormal samples. Along with that, with 20% of TSS, the XAIID-SCPS technique identifies 13,343 normal and 11,578 abnormal samples. Finally, with 30% of TSS, the XAIID-SCPS technique identifies 19,787 normal and 17,364 abnormal samples.

In Table 2, detailed intrusion recognition outcomes of the XAIID-SCPS method on the NSLKDD2015 dataset are provided under 80:20 of TRS/TSS. The experimental results demonstrated that the proposed model properly recognizes normal and anomaly samples under 80:20 of TRS/TSS. With 80% of TRS, the XAIID-SCPS technique obtains an average $accu_{bal}$ of 98.86%, $prec_n$ of 98.93%, $reca_l$ of 98.86%, F_{score} of 98.89%, and AUC_{score} of 98.86%. Meanwhile, with 20% of TSS, the XAIID-SCPS methodology acquires an average $accu_{bal}$ of 98.87%, $prec_n$ of 98.95%, $reca_l$ of 98.87%, F_{score} of 98.91%, and AUC_{score} of 98.87%. Additionally, for 30% of TSS, the XAIID-SCPS method obtains an average $accu_{bal}$ of 98.32%, $prec_n$ of 98.28%, $reca_l$ of 98.32%, F_{score} of 98.30%, and AUC_{score} of 98.32%.

Table 2. Intrusion recognition outcome of XAIID-SCPS system on the NSLKDD2015 dataset.

NSLKDD 2015 Dataset					
Class	$Accu_{bal}$	$Prec_n$	$Reca_l$	F_{score}	AUC_{score}
Training Phase (80%)					
Normal	99.40	98.55	99.40	98.97	98.86
Anomaly	98.31	99.30	98.31	98.81	98.86
Average	**98.86**	**98.93**	**98.86**	**98.89**	**98.86**
Testing Phase (20%)					
Normal	99.51	98.47	99.51	98.98	98.87
Anomaly	98.24	99.43	98.24	98.83	98.87
Average	**98.87**	**98.95**	**98.87**	**98.91**	**98.87**
Training Phase (70%)					
Normal	98.04	98.70	98.04	98.37	98.28
Anomaly	98.52	97.76	98.52	98.14	98.28
Average	**98.28**	**98.23**	**98.28**	**98.25**	**98.28**
Testing Phase (30%)					
Normal	98.11	98.70	98.11	98.41	98.32
Anomaly	98.52	97.85	98.52	98.19	98.32
Average	**98.32**	**98.28**	**98.32**	**98.30**	**98.32**

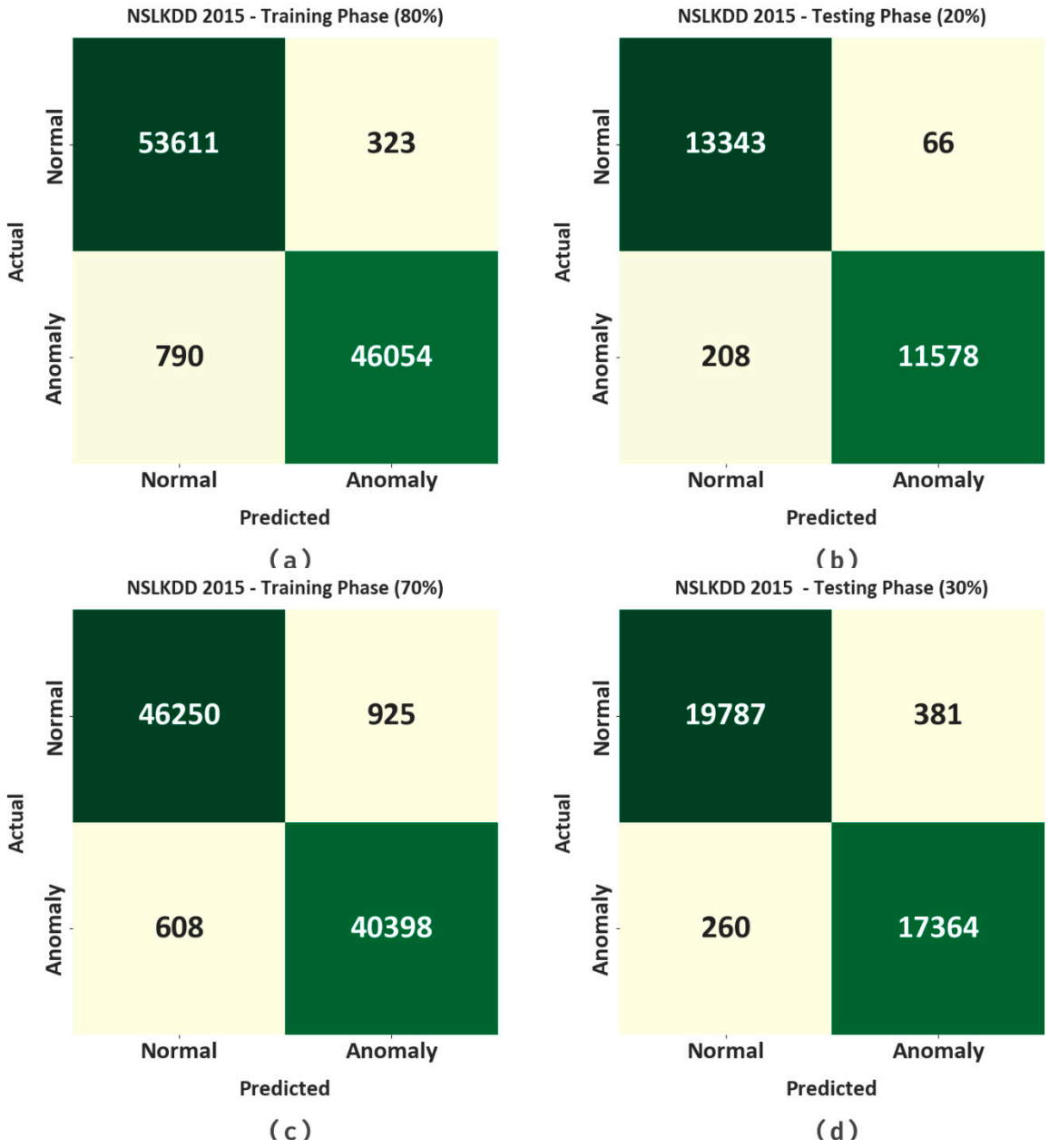

Figure 3. Confusion matrices of XAIID-SCPS system on NSLKDD2015 dataset (**a**,**b**) TRS/TSS of 80:20 and (**c**,**d**) TRS/TSS of 70:30.

The TACY and VACY of the XAIID-SCPS model on the NSLKDD2015 dataset are represented in Figure 4. The figure designated the XAIID-SCPS model has shown enhanced performance with maximal values of TACY and VACY. It is visible that the XAIID-SCPS model has maximum TACY results.

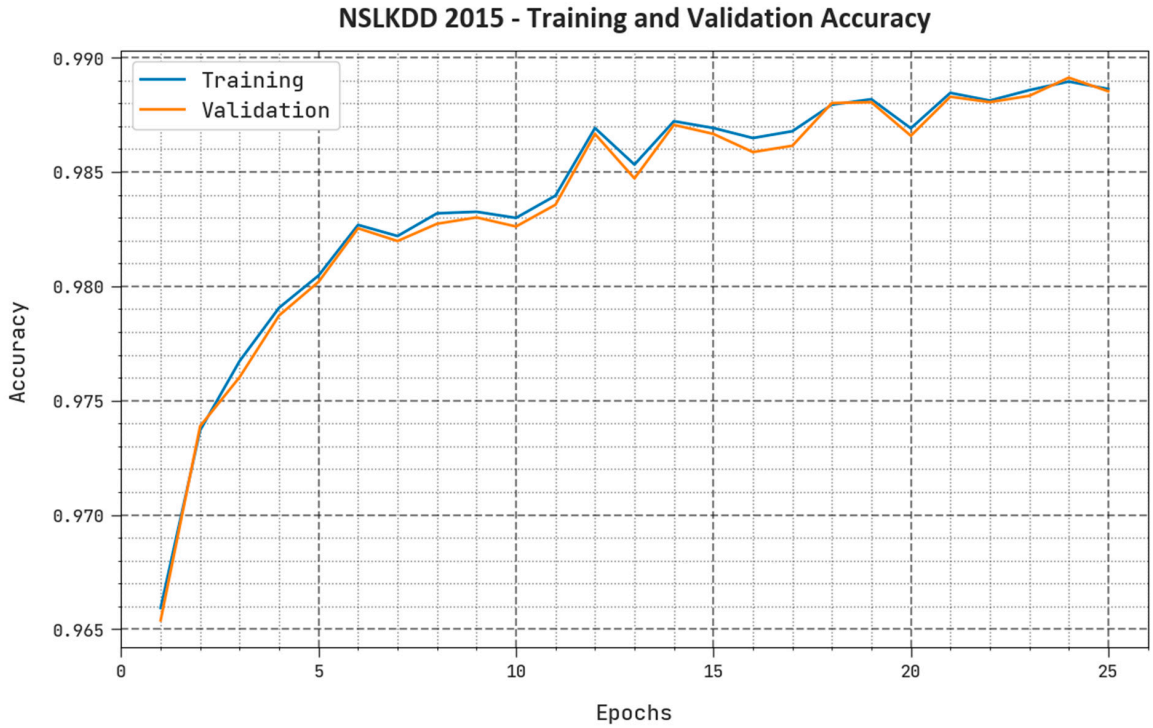

Figure 4. TACY and VACY outcome of XAIID-SCPS system on NSLKDD2015 dataset.

The TLOS and VLOS of the XAIID-SCPS model on the NSLKDD2015 dataset are represented in Figure 5. The figure inferred that the XAIID-SCPS model has superior performance with least values of TLOS and VLOS. It is visible that the XAIID-SCPS model has resulted in reduced VLOS outcomes.

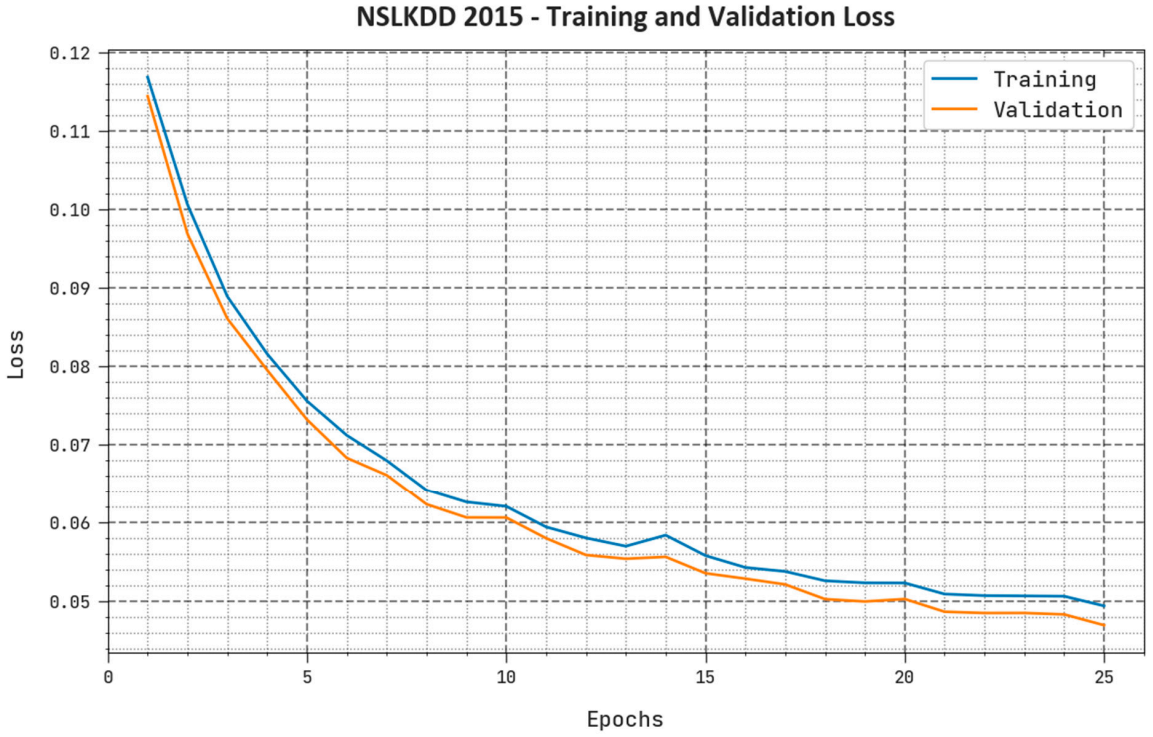

Figure 5. TLOS and VLOS outcome of XAIID-SCPS system on NSLKDD2015 dataset.

The confusion matrices of the XAIID-SCPS methodology in the CICIDS 2017 dataset are shown in Figure 6. The results recognized the normal and abnormal samples proficiently. For instance, with 80% of TRS, the XAIID-SCPS technique identifies 39,508 normal and 39,422 abnormal samples. Along with that, with 20% of TSS, the XAIID-SCPS technique identifies 9774 normal and 9950 abnormal samples. Finally, with 30% of TSS, the XAIID-SCPS technique identifies 14,582 normal and 14,505 abnormal samples.

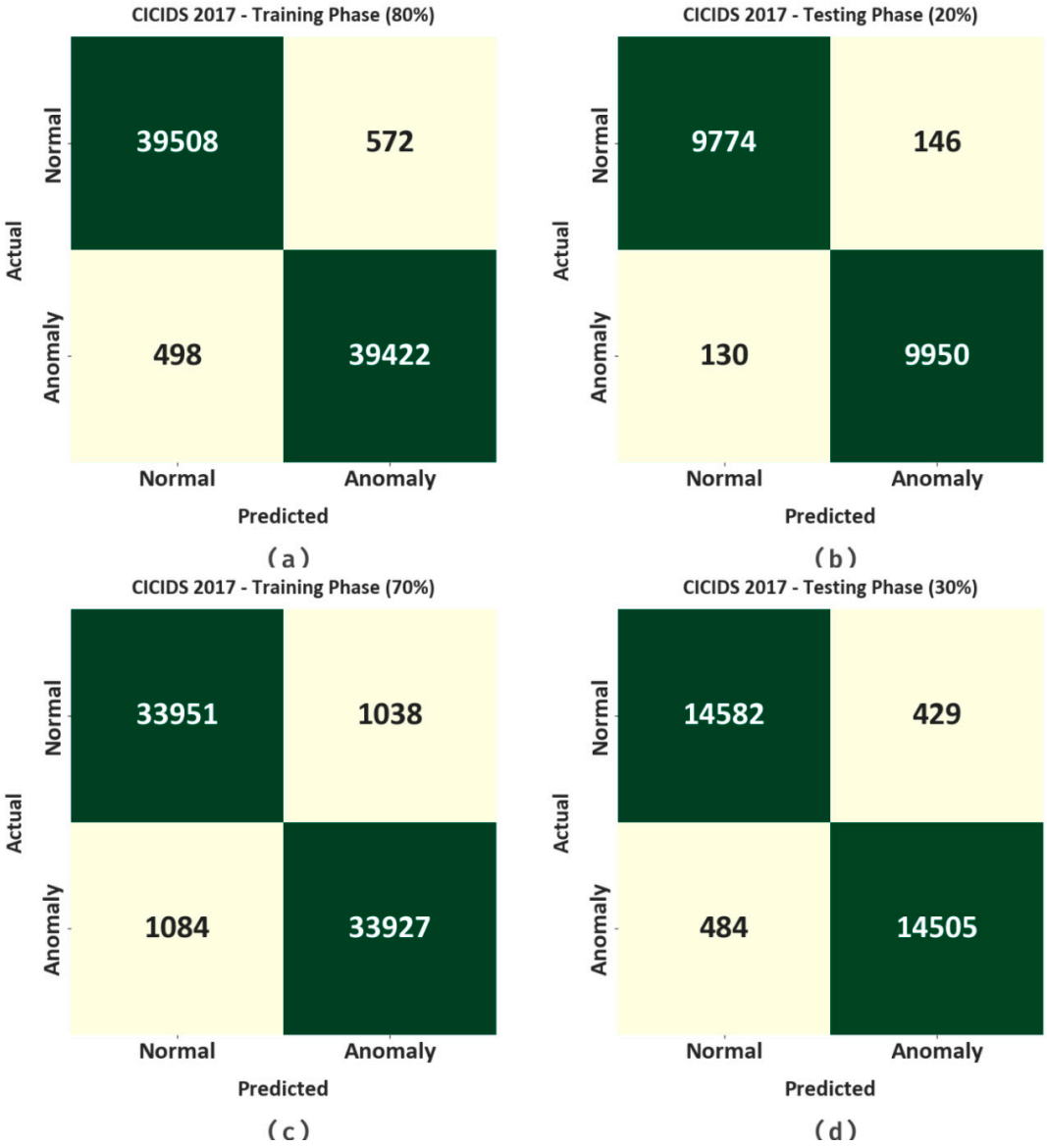

Figure 6. Confusion matrices of XAIID-SCPS system on CICIDS 2017 dataset (**a**,**b**) TRS/TSS of 80:20 and (**c**,**d**) TRS/TSS of 70:30.

In Table 3, a brief intrusion recognition outcome of the XAIID-SCPS method on the CICIDS 2017 dataset is provided under 70:30 of TRS/TSS. The experimental results demonstrated that the proposed model properly recognizes normal and anomaly samples under 70:30 of TRS/TSS. With 80% of TRS, the XAIID-SCPS technique obtains an average $accu_{bal}$ of 98.66%, $prec_n$ of 98.66%, $reca_l$ of 98.66%, F_{score} of 98.66%, and AUC_{score} of 98.66%. Meanwhile, with 20% of TSS, the XAIID-SCPS technique obtains an average $accu_{bal}$ of 98.62%, $prec_n$ of 98.62%, $reca_l$ of 98.62%, F_{score} of 98.62%, and AUC_{score} of 98.62%. Moreover, for 30% of TSS, the XAIID-SCPS technique obtains an average $accu_{bal}$ of 96.96%, $prec_n$ of 96.96%, $reca_l$ of 96.96%, F_{score} of 96.96%, and AUC_{score} of 96.96%.

The TACY and VACY of the XAIID-SCPS model on the CICIDS 2017 database are represented in Figure 7. The figure shows that the XAIID-SCPS method has shown enhanced performance with increased values of TACY and VACY. It is visible that the XAIID-SCPS model has higher TACY outcomes.

Table 3. Intrusion recognition outcome of XAIID-SCPS system on the CICIDS 2017 dataset.

CICIDS 2017 Dataset					
Class	$Accu_{bal}$	$Prec_n$	$Reca_l$	F_{score}	AUC_{score}
Training Phase (80%)					
Normal	98.57	98.76	98.57	98.66	98.66
Anomaly	98.75	98.57	98.75	98.66	98.66
Average	**98.66**	**98.66**	**98.66**	**98.66**	**98.66**
Testing Phase (20%)					
Normal	98.53	98.69	98.53	98.61	98.62
Anomaly	98.71	98.55	98.71	98.63	98.62
Average	**98.62**	**98.62**	**98.62**	**98.62**	**98.62**
Training Phase (70%)					
Normal	97.03	96.91	97.03	96.97	96.97
Anomaly	96.90	97.03	96.90	96.97	96.97
Average	**96.97**	**96.97**	**96.97**	**96.97**	**96.97**
Testing Phase (30%)					
Normal	97.14	96.79	97.14	96.96	96.96
Anomaly	96.77	97.13	96.77	96.95	96.96
Average	**96.96**	**96.96**	**96.96**	**96.96**	**96.96**

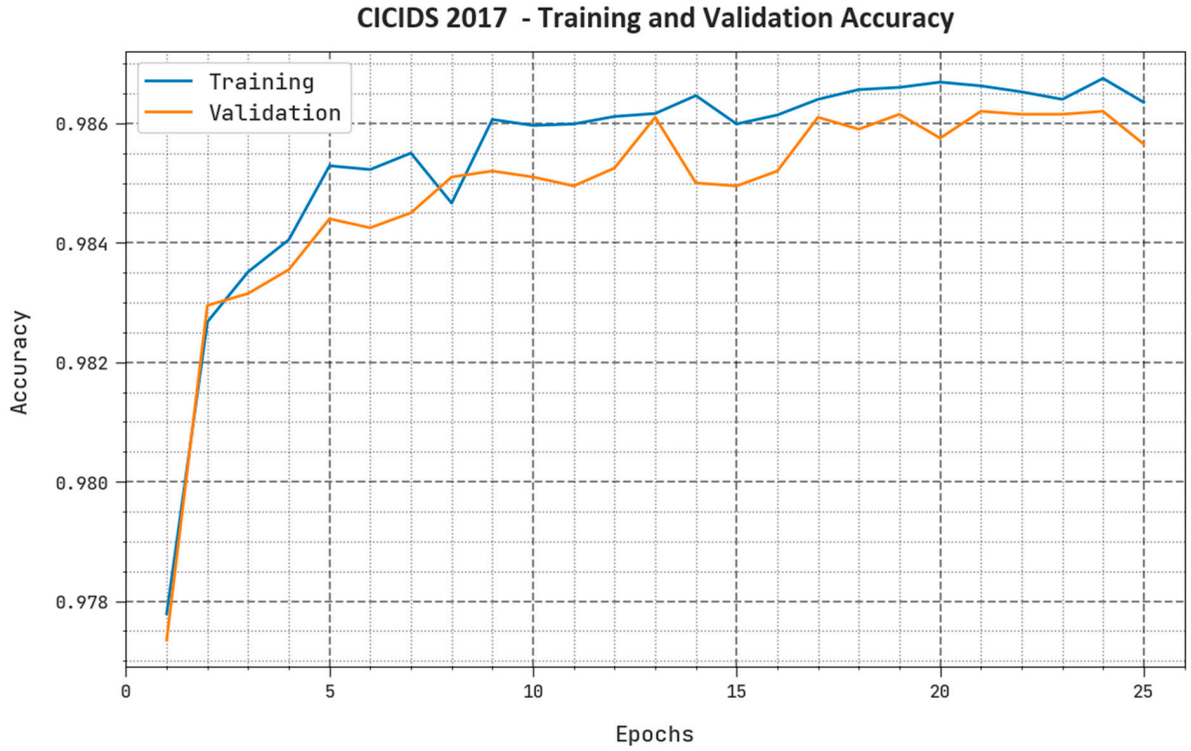

Figure 7. TACY and VACY outcome of XAIID-SCPS system on CICIDS 2017 dataset.

The TLOS and VLOS of the XAIID-SCPS model on the CICIDS 2017 dataset are represented in Figure 8. The figure indicated that the XAIID-SCPS model has better performance with the least values of TLOS and VLOS. It is visible that the XAIID-SCPS model has resulted in reduced VLOS outcomes.

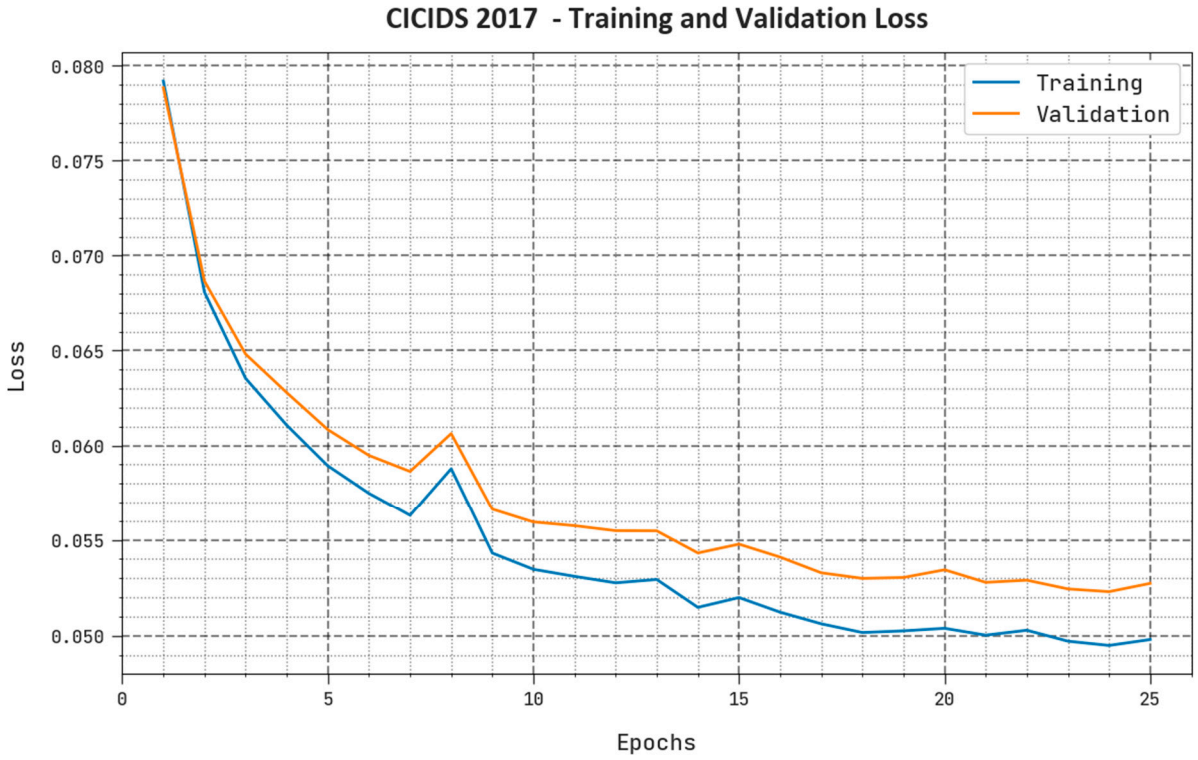

Figure 8. TLOS and VLOS outcome of XAIID-SCPS system on CICIDS 2017 dataset.

In Table 4 and Figure 9, comparative results of the XAIID-SCPS technique with existing models take place [27]. The results indicate that the XAIID-SCPS method shows maximum performance over other existing models. Based on $accu_y$, the XAIID-SCPS technique results in an increasing $accu_y$ of 98.87% while the FURIA, AE-RF, Forest-PA, WIZARD, GSAE, and LIB-SVM models reach a reducing $accu_y$ of 98.14%, 97.62%, 96.72%, 96.64%, 97.63%, and 96.57%, respectively.

Table 4. Comparative outcome of XAIID-SCPS system with other systems.

Methods	Accuracy	Precision	Recall	F1-Score
XAIID-SCPS	98.87	98.95	98.87	98.91
FURIA	98.14	97.57	96.93	98.26
AE-RF	97.62	97.35	97.79	97.30
Forest-PA	96.72	96.97	97.32	98.13
WISARD	96.64	97.58	97.29	98.65
GSAE	97.63	95.97	98.39	98.19
LIB-SVM	96.57	96.96	96.83	97.92

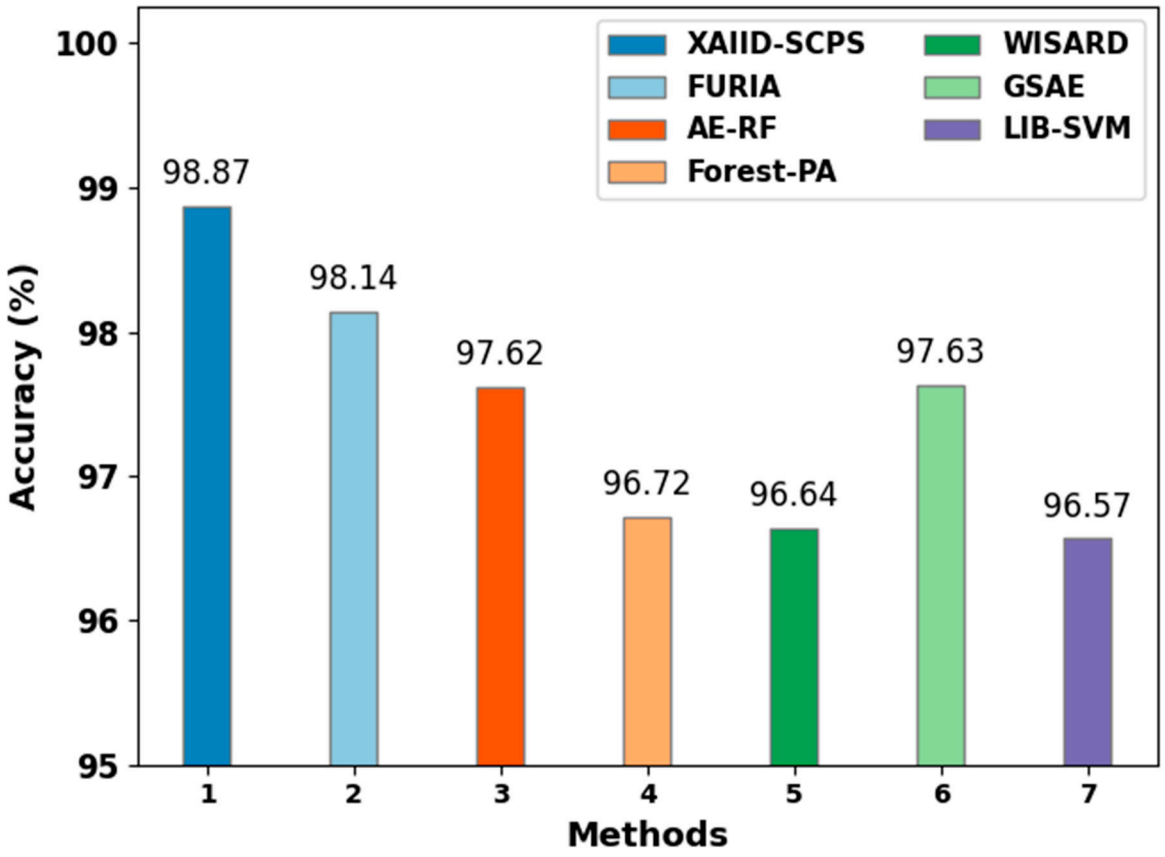

Figure 9. $Accu_y$ outcome of XAIID-SCPS system with other systems.

Meanwhile, based on $prec_n$, the XAIID-SCPS method results in an increasing $prec_n$ of 98.95% while the FURIA, AE-RF, Forest-PA, WISARD, GSAE, and LIB-SVM models reach a reducing $prec_n$ of 97.57%, 97.35%, 96.97%, 97.58%, 95.97%, and 96.96%, respectively. Eventually, based on F_{score}, the XAIID-SCPS technique results in an increasing F_{score} of 98.91% while the FURIA, AE-RF, Forest-PA, WISARD, GSAE, and LIB-SVM models reach a reducing F_{score} of 98.14%, 98.26%, 97.30%, 98.13%, 98.65%, 98.19%, and 97.92%, correspondingly. These results highlighted the betterment of the XAIID-SCPS technique for intrusion detection purposes.

5. Conclusions

In this study, we have presented an automated intrusion detection technique named XAIID-SCPS technique for the CPS platform. The presented XAIID-SCPS approach primarily concentrates on the detection and classification of intrusions in the CPS platform. In the XAIID-SCPS technique, several subprocesses are involved, namely, data pre-processing, HEGSO-based feature selection, IENN-based classification, and EFFO-based parameter tuning. Moreover, the XAIID-SCPS technique integrates the XAI approach LIME for better understanding and explainability of the black-box method for accurate classification of intrusions. The simulation values of the XAIID-SCPS technique are tested on a benchmark intrusion dataset and the outcomes prove the promising performance of the XAIID-SCPS technique over other recent approaches. In the future, data clustering and outlier removal methodologies can be designed to enrich the detection performance of the XAIID-SCPS technique. Moreover, the proposed model can be extended to the design of ensemble voting classifiers to improve the detection rate of the XAIID-SCPS technique.

Author Contributions: Conceptualization, L.A.; Methodology, M.S.M.; Software, M.A. and A.A.A.; Validation, M.A., H.M. and M.A.H.; Formal analysis, A.A.A.; Investigation, M.A.; Resources, H.M.; Data curation, H.M.; Writing–original draft, L.A., M.S.M., H.M. and M.A.H.; Writing–review & editing, M.S.M., M.A., M.A.H. and A.A.A.; Visualization, A.A.A.; Supervision, L.A.; Project administration, M.A.H.; Funding acquisition, L.A. All authors have read and agreed to the published version of the manuscript.

Funding: The authors extend their appreciation to the Deanship of Scientific Research at King Khalid University for funding this work through Large Groups Project under grant number (134/44). Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2023R349), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. Research Supporting Project number (RSP2023R787), King Saud University, Riyadh, Saudi Arabia. This study is supported via funding from Prince Sattam bin Abdulaziz University project number (PSAU/2023/R/1444).

Institutional Review Board Statement: This article does not contain any studies with human participants performed by any of the authors.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data sharing is not applicable to this article as no datasets were generated during the current study.

Conflicts of Interest: The authors declare that they have no conflict of interest. The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

References

1. Arisdakessian, S.; Wahab, O.A.; Mourad, A.; Otrok, H.; Guizani, M. A survey on iot intrusion detection: Federated learning, game theory, social psychology and explainable ai as future directions. *IEEE Internet Things J.* **2022**. [CrossRef]
2. Capuano, N.; Fenza, G.; Loia, V.; Stanzione, C. Explainable Artificial Intelligence in CyberSecurity: A Survey. *IEEE Access* **2022**, *10*, 93575–93600. [CrossRef]
3. Khakpour, N. Security Explainability Challenges in Cyber-Physical Systems. In *Explainable Software for Cyber-Physical Systems (ES4CPS)*; Gesellschaft für Informatik: Bonn, Germany, 2019; p. 44.
4. Zhang, Z.; Hamadi, H.A.; Damiani, E.; Yeun, C.Y.; Taher, F. Explainable Artificial Intelligence Applications in Cyber Security: State-of-the-Art in Research. *arXiv* **2022**, arXiv:2208.14937. [CrossRef]
5. Kabir, M.H.; Hasan, K.F.; Hasan, M.K.; Ansari, K. Explainable Artificial Intelligence for Smart City Application: A Secure and Trusted Platform. In *Explainable Artificial Intelligence for Cyber Security*; Springer: Cham, Switzerland, 2022; pp. 241–263.
6. Khanapuri, E.; Chintalapati, T.; Sharma, R.; Gerdes, R. Learning-based adversarial agent detection and identification in cyber physical systems applied to autonomous vehicular platoon. In Proceedings of the 2019 IEEE/ACM 5th International Workshop on Software Engineering for Smart Cyber-Physical Systems (SEsCPS), Montreal, QC, Canada, 28 May 2019; pp. 39–45.
7. Panigrahi, R.; Borah, S.; Pramanik, M.; Bhoi, A.K.; Barsocchi, P.; Nayak, S.R.; Alnumay, W. Intrusion detection in cyber–physical environment using hybrid Naïve Bayes—Decision table and multi-objective evolutionary feature selection. *Comput. Commun.* **2022**, *188*, 133–144. [CrossRef]
8. Amarasinghe, K.; Wickramasinghe, C.; Marino, D.; Rieger, C.; Manicl, M. Framework for data driven health monitoring of cyber-physical systems. In Proceedings of the 2018 Resilience Week (RWS), Denver, CO, USA, 20–23 August 2018; pp. 25–30.
9. Radanliev, P.; De Roure, D.; Page, K.; Van Kleek, M.; Santos, O.; Maddox, L.T.; Burnap, P.; Anthi, E.; Maple, C. Design of a dynamic and self-adapting system, supported with artificial intelligence, machine learning and real-time intelligence for predictive cyber risk analytics in extreme environments–cyber risk in the colonisation of Mars. *Saf. Extrem. Environ.* **2020**, *2*, 219–230. [CrossRef]
10. Radanliev, P.; De Roure, D.; Walton, R.; Van Kleek, M.; Montalvo, R.M.; Maddox, L.T.; Santos, O.; Burnap, P.; Anthi, E. Artificial intelligence and machine learning in dynamic cyber risk analytics at the edge. *SN Appl. Sci.* **2020**, *2*, 1–8. [CrossRef]
11. Munir, M.; Dipro, S.H.; Hasan, K.; Islam, T.; Shetty, S. Artificial Intelligence-Enabled Exploratory Cyber-Physical Safety Analyzer Framework for Civilian Urban Air Mobility. *Appl. Sci.* **2023**, *13*, 755. [CrossRef]
12. Colelli, R.; Magri, F.; Panzieri, S.; Pascucci, F. Anomaly-Based Intrusion Detection System for Cyber-Physical System Security. In Proceedings of the 2021 29th Mediterranean Conference on Control and Automation (MED), Bari, Italy, 22–25 June 2021; pp. 428–434.
13. Huang, K.; Zhou, C.; Tian, Y.C.; Yang, S.; Qin, Y. Assessing the physical impact of cyberattacks on industrial cyber-physical systems. *IEEE Trans. Ind. Electron.* **2018**, *65*, 8153–8162.
14. Schneider, P.; Böttinger, K. High-performance unsupervised anomaly detection for cyber-physical system networks. In Proceedings of the 2018 Workshop on Cyber-Physical Systems Security and Privacy, Toronto, ON, Canada, 15–19 October 2018; pp. 1–12.

15. Sharma, V.; You, I.; Yim, K.; Chen, R.; Cho, J.H. BRIoT: Behavior rule specification-based misbehavior detection for IoT-embedded cyber-physical systems. *IEEE Access* **2019**, *7*, 118556–118580. [CrossRef]
16. Huang, K.; Zhou, C.; Qin, Y.; Tu, W. A game-theoretic approach to cross-layer security decision-making in industrial cyber-physical systems. *IEEE Trans. Ind. Electron.* **2019**, *67*, 2371–2379. [CrossRef]
17. Huang, L.; Zhu, Q. A dynamic games approach to proactive defense strategies against advanced persistent threats in cyber-physical systems. *Comput. Secur.* **2020**, *89*, 101660. [CrossRef]
18. Wang, Z.; Li, Z.; He, D.; Chan, S. A lightweight approach for network intrusion detection in industrial cyber-physical systems based on knowledge distillation and deep metric learning. *Expert Syst. Appl.* **2022**, *206*, 117671. [CrossRef]
19. Tang, B.; Lu, Y.; Li, Q.; Bai, Y.; Yu, J.; Yu, X. A Diffusion Model Based on Network Intrusion Detection Method for Industrial Cyber-Physical Systems. *Sensors* **2023**, *23*, 1141. [CrossRef]
20. Ramadevi, P.; Baluprithviraj, K.N.; Pillai, V.A.; Subramaniam, K. Deep Learning Based Distributed Intrusion Detection in Secure Cyber Physical Systems. *Intell. Autom. Soft Comput.* **2022**, *34*, 2067–2081. [CrossRef]
21. Alohali, M.A.; Al-Wesabi, F.N.; Hilal, A.M.; Goel, S.; Gupta, D.; Khanna, A. Artificial intelligence enabled intrusion detection systems for cognitive cyber-physical systems in industry 4.0 environment. *Cogn. Neurodynamics* **2022**, *16*, 1045–1057. [CrossRef] [PubMed]
22. Dutta, A.K.; Negi, R.; Shukla, S.K. Robust multivariate anomaly-based intrusion detection system for cyber-physical systems. In *Cyber Security Cryptography and Machine Learning: 5th International Symposium, CSCML 2021, Be'er Sheva, Israel, 8–9 July 2021, Proceedings 5*; Springer International Publishing: Berlin/Heidelberg, Germany, 2021; pp. 86–93.
23. Upadhyay, P.; Marriboina, V.; Kumar, S.; Kumar, S.; Shah, M.A. An Enhanced Hybrid Glowworm Swarm Optimization Algorithm for Traffic-Aware Vehicular Networks. *IEEE Access* **2022**, *10*, 110136–110148. [CrossRef]
24. Zhang, J.; Ding, X.; Hu, D.; Guo, B.; Jiang, Y. Performance Evaluation of Enterprise Collaboration Based on an Improved ENN and AHP-EW. *Appl. Sci.* **2022**, *12*, 5941. [CrossRef]
25. Sun, H.; Li, W.; Zheng, L.; Ling, S.; Fu, W. Adaptive co-simulation method and platform application of drive mechanism based on Fruit Fly Optimization Algorithm. *Prog. Nucl. Energy* **2022**, *153*, 104397. [CrossRef]
26. Zafar, M.R.; Khan, N. Deterministic local interpretable model-agnostic explanations for stable explainability. *Mach. Learn. Knowl. Extr.* **2021**, *3*, 525–541. [CrossRef]
27. Duhayyim, M.A.; Alissa, K.A.; Alrayes, F.S.; Alotaibi, S.S.; Tag El Din, E.M.; Abdelmageed, A.A.; Yaseen, I.; Motwakel, A. Evolutionary-Based Deep Stacked Autoencoder for Intrusion Detection in a Cloud-Based Cyber-Physical System. *Appl. Sci.* **2022**, *12*, 6875. [CrossRef]