Preprints.org

Article

Not peer-reviewed version

Federated XAI IDS: An Explainable and Safeguarding privacy Approach to Detect Intrusion Combining Federated Learning and SHAP

Kazi Fatema *, Samrat Kumar Dey, Mehrin Anannya, Risala Tahsin Khan, Mohammad Rashid, SU Chunhua *, Rashed Mazumder *

Posted Date: 26 March 2025

doi: 10.20944/preprints202503.1902.v1

Keywords: Cyber Security; FedXAIIDS(Federated Explainable IDS); Intrusion Detection System(IDS); XAI(Explainable AI); (SHAP)SHapley Additive Explanaition; ANN

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2025 doi:10.20944/preprints202503.1902.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Federated XAI IDS: An Explainable and Safeguarding Privacy Approach to Detect Intrusion Combining Federated Learning and SHAP

Kazi Fatema [1,*], Samrat Kumar Dey [2], Mehrin Anannya [2], Risala Tahsin Khan [1], Mohammad Mamunur Rashid [2], SU Chunhua [3,*] and Rashed Mazumder [1,*]

[1] Institute of Information Technology, Jahangirnagar University, Dhaka-1342
[2] School of Science & Technology, Bangladesh Open University, Dhaka
[3] Department of Computer Science and Engineering, University of Aizu, Japan
* Correspondence: kazifatema1975@gmail.com (K.F.); chsu@u-aizu.ac.jp (S.C.); rakhu345@yahoo.com (R.M.)

Abstract: Intrusion Detection Systems (IDS) are crucial module of cybersecurity which is designed to identify unauthorized activities in network environments. Traditional IDS, on the other hand, have a number of problems, such as high rates of inaccurate positives and inaccurate negatives and a lack of explainability that makes it difficult to provide adequate protection. Furthermore, centralized IDS approaches have issues with interpretability and data protection, especially when dealing with sensitive data. In order to overcome these drawbacks, we provide Federated XAI IDS, a brand-new explainable and privacy-preserving IDS that improves security and interpretability by fusing Federated Learning (FL) with Shapley Additive Explanations (SHAP). Our approach enables IDS models to be collaboratively trained across multiple decentralized devices while ensuring that local data remains securely on edge nodes, thus mitigating privacy risks. The Artificial Neural Network (ANN)-based IDS is distributed across four clients in a federated setup using the CICIoT2023 dataset, with model aggregation performed via FedAvg. The proposed method demonstrated efficacy in intrusion detection, achieving 88.4% training and 88.2% testing accuracy. Furthermore, SHAP was utilized to analyze feature importance, providing a deeper comprehension of the critical attributes influencing model predictions. Transparency is improved and the model becomes more dependable and interpretable thanks to the feature importance ranking that SHAP produces. Our findings demonstrate how well Federated XAI IDS handles the two problems of explainability and privacy in intrusion detection. This dissertation accelerates the major establishment in the creation of safe, interpretable, and decentralized intrusion detection systems (IDS) for contemporary cybersecurity applications by utilizing federated learning and explainable AI (XAI).

Keywords: Cyber Security; FedXAIIDS(Federated Explainable IDS); Intrusion Detection System(IDS); XAI(Explainable AI); (SHAP)SHapley Additive Explanaition; ANN

1. Introduction

In the digital age, cybersecurity — the practice of safeguarding frameworks, networks, and confidential data from unwelcome inspection, distress, emerging cyber hazards as well [1], has become the essential battle line. Since people, organizations, and governments grow more interconnected and dependent on digital infrastructure, it is crucial to guarantee data availability, confidentiality, and integrity. In order to safeguard their valuable assets and lower risks, companies need to proactively create multilayer protection measures in response to evolving cyber threats. Within the ever-changing field of cybersecurity, safeguarding private data from a continuously increasing spectrum of cyber attacks is essential. In order to discover and prevent undesirable activity in network traffic, intrusion detection systems (IDS) are essential. They are the first line of defense against malicious attacks. An intrusion detection system (IDS) scans over malicious activity or unauthorized access through

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2025 doi:10.20944/preprints202503.1902.v1

analysis of traffic dynamics, applications and session behavior, and signature-based features [2]. By emphasizing anomalies and recognized attack patterns, IDS alerts help organizations promptly address possible security risks.

In this new age of interconnected world, approximately 50% of startups disclose experiencing information theft, underscoring the urgency of robust IDS solutions in institutional security frameworks [3]. The integration of Deep Learning (DL), Machine Learning (ML) along with IDS has acquired substantial traction due to their superior classification accuracy. Nonetheless, conventional machine learning-based intrusion detection systems encounter considerable privacy and security vulnerabilities owing to their dependence on centralized data storage and transmission [4]. Additionally, The challenges of developing effective IDS systems include the substantial quantity of network information and the prevalence of imbalanced datasets, where minor yet critical attack types are often underrepresented [5].

Higher performance in interpreting complicated patterns within network data is made possible by recent developments in deep learning, especially the adoption of transformer models, which have showed significant prospects in sequencing frameworks and anomaly discovery[6]. Intrusion Detection Systems (IDS), which are crucial for locating and preventing existing network intrusions, are particularly affected by this.

However, there are serious concerns regarding confidentiality, with the use of deep learning-based IDS frameworks, especially when models are trained using private network traffic data. These issues can be effectively addressed by Federated Learning (FL), which permits for decentralized coordinated infrastructure refinement while maintaining the privacy and security of confidential information [7]. This research introduces a federated learning-based intrusion detection system that employs artificial neural networks as the local-end prediction system and incorporates SHapley Additive Explanations in order to enhance comprehensibility The decentralized training approach guarantees confidentiality while facilitating visibility in detecting threats to networks. This piece of literature is the expanded version of the initial manuscript previously laid out in[8]. Explainable artificial intelligence (XAI) has significantly enhanced intrusion detection systems (IDS) by providing accountability, understanding, and integrity in decision-making. Traditional IDS models, particularly those based on deep learning, frequently function as "black boxes," complicating security analysts' understanding of their predictions. SHapley Additive ExPlanations (SHAP) has emerged as a powerful XAI approach, offering feature attribution scores that elucidate the influence of different input characteristics on the conclusions of the models. Employing SHAP in Intrusion Detection Systems aids researchers and practitioners in identifying significant attack patterns, reducing false positives, and enhancing model robustness. Research, including [9], demonstrates that SHAP provides equitable and consistent feature importance ratings, making it an ideal method for enhancing the interpretability of IDS. Moreover, SHAP facilitates feature selection and optimization, hence enhancing generalization and efficiency in IDSs that are constructed deploying FL, where privacy preservation is paramount [10]. Utilizing SHAP in FL-based IDS confidentiality security insights while safeguarding raw data, hence maintaining user privacy and delivering actionable knowledge for cyber defense.

The CICIoT2023 dataset[11], a real-world dataset with contemporary attack types, has been used for the experiments. The system's performance and generalizability in a range of intrusion scenarios, as well as its capacity to identify known and new network threats, has been the main areas of focus. In addition, the integration of explainable artificial intelligence (XAI) techniques, particularly SHAP, will provide essential transparency within the IDS framework. This comprehensibility constitutes an essential towards instilling assurance in AI-driven frameworks, specifically when legal and regulatory compliance is a factor in the deployment of network security solutions. The new study builds on the foundations laid by focusing on explainable AI and federated learning, while the earlier work concentrated on these conventional approaches. The goal is to improve the privacy, efficacy, and interpretability of IDS solutions. This research builds upon prior work with the CIC_IDS_2018 dataset, where the challenge of imbalanced datasets was addressed by developing an IDS using

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2025 doi:10.20944/preprints202503.1902.v1

balanced data and exploring unorthodox categorization methods incorporating the Dendritic Cell Algorithm (DCA) [12]. In order to improve transparency and interpretability and facilitate a more thorough comprehension of model decisions, explainable artificial intelligence (XAI) approaches must be incorporated into the IDS framework. By elucidating the reasoning behind certain results, this transparency contributes to the development of trust in AI-driven frameworks, which is crucial for fulfilling legal and regulatory obligations.

This research attempts to tackle important network security issues by designing and deploying a sophisticated Intrusion Detection System (IDS) that integrates Explainable Artificial Intelligence (XAI), Federated Learning (FL), and Machine Learning (ML). This study aims to achieve the following goals:

1. To assess the efficacy of an Intrusion Detection System utilizing cutting-edge Machine Learning methodologies, Federated Learning for decentralized training, and Explainable Artificial Intelligence for enhanced interpretability and transparency.
2. To minimize computational overhead by utilizing Federated Learning, which allows decentralized data processing and eliminates the need for central data aggregation.
3. To reduce the dangers of single points of failure and centralized data breaches by implementing a distributed, node-based architecture in Federated Learning.
4. To ensure the reliability and trustworthiness of IDS predictions by employing Explainable AI, which provides insights into model decisions and fosters greater user trust in automated systems.

This research work has been structured in the following manner: Section-2 highlights relevant research on XAI, Federated Learning, and IDS. Section-3 details the proposed methodology, including model architecture and SHAP-based interpretability. Section-4 discusses the experimental setup, along with results and analysis. Section-5 discusses key findings, and Section-6 finishes with contributions and future study directions.

2. Literature Review

This sections analyzes a number of significant research works that paved the way for the proposed framework. The growing popularity popularity of Internet of Things(IoT) networks has generated considerable security concerns, prompting the creation of effective IDSs. Recent improvements in FL represents a viable approach, facilitating decentralized learning while safeguarding data privacy. This section examines current research initiatives that combine FL with DL.

Table 1. A tabular representation of the FL-based infrastructure and its corresponding challenges.

Authors	FL technique used (Yes/No)	Framework
Sinh-Ngoc et al. [13]	No	Employed CNN architecture for the categorization.
Kanimozhi et al. [7].	No	The applied classifiers for detecting network assaults include NB, K-Nearest Neighbor, RF, Adaboost with Decision Tree, SVM, and ANN. These classifiers specifically target the detection of Botnet network assaults.
Bertoli et al. [14]	Yes	Constructed a multilayered autonomous FL architecture that integrates an autoencoder with an energy flow classifier, enabling enhanced feature extraction and classification performance while maintaining privacy in a distributed learning environment.
Toldinas et al. [15]	No	The initial processing technique that combines a predetermined number of network flow feature records. Three independent ML methodologies: , Federated transfer learning, Traditional transfer learning, and Federated learning were used on NIDS employing deep learning for image classification.
Markovic et al. [16]	Yes	Implemented a Federated Learning (FL) model that utilizes the shared model incorporating RF, enabling learning across multiple consumers collaboratively while safeguarding the privacy of informations.
Lazzarini et al. [17].	Yes	Developed a IDS incorporating FL, a shallow ANN as the regional framework and FedAvg as the aggregation method.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2025 doi:10.20944/preprints202503.1902.v1

Torre et al. [18] initiated an IDS built on FL adopting a 1D Convolutional Neural Network (CNN) to secure IoT networks. This framework enhances privacy through Diffie–Hellman Key Exchange, Dynamic Security, and Homomorphic Cryptography. Almadhor et al. [19] introduced a Federated Deep Neural Network (FDNN) for detecting and preventing Distributed Denial-of-Service (DDoS) integrating Explainable Artificial Intelligence (XAI) in IoT networks. Their methodology employed FDNNs trained across three client devices over multiple rounds without sharing raw data. Furthermore, XGBoost was used with SHapley Additive ExPlanations (SHAP) for selecting features, which improved model comprehensibility This method successfully preserved robustness, scalability, and confidentiality while obtaining high detection accuracy. To further optimize FL-based IDSs, Alsaleh et al. [20] offered a semi-distributed FL model that clusters IoT devices and assigns a cluster head to reduce communication overhead. The model incorporated Bidirectional LSTM (BiLSTM), Long Short-Term Memory (LSTM), and Wasserstein Generative Adversarial Networks (WGAN) to enhance intrusion detection, particularly focusing on DDoS attacks in scenarios with scarce resources. Their evaluation implementing CICIoT2023 dataset revealed BiLSTM as the most efficient model due to its optimized size. Testing on, WUSTL-IIoT-2021, Edge-IIoTset, and BoT-IoT, further confirmed its superior detection accuracy. An unorthodox Secure and Authenticated structure built on Federated Learning-employing Blockchain (SA-FLIDS) was initiated by Bensaid et al. [21] to enhance security in advanced healthcare systems that are enabled by fog-IoMT. Experimental results using CICIoT2023 and EdgeIIoTset datasets demonstrated the framework's strong resilience against adversarial attacks while preserving confidentiality of the data and deducting the expenses of communication. Sun et al. [22] addressed the dilemma of attack class dispersion in FL-based IDSs by proposing FedMADE, an adaptive collaborative framework. FedMADE groups IoT gadgets based on the trends on the traffic and incorporates local approaches according to their significance to entire evaluation. This approach significantly improved the accuracy of minority attacks classification to 71.07% compared to existing FL methods for non-IID data. Additionally, FedMADE exhibited robustness against poisoning attacks while incurring only a 4.7% latency overhead (5.03 seconds per iteration) contrasting with FedAvg, without uploading computational weigh on IoT gadgets. These research works collectively underlines the efficacy of FL-based IDSs in intensifying cybersecurity within IoT networks. The combination of deep learning simulations, privacy-preserving techniques, and blockchain technology shows great promise for protecting modern IoT systems. However, issues such as communication overhead, data disparity, and adversarial resilience remain critical areas for further research and optimization.

Deep Learning (DL) and Machine Learning (ML) have become foundational approaches for constructing IDS, significantly improving network confidentiality by spotting various attacks, anomalies as well. Several ML techniques are widely used for their robustness and efficiency. Support Vector Machine (SVM) is valued for handling high-dimensional data, while Decision Tree (DT) offers simplicity and interpretability. Random Forest (RF), an association of decision trees, amplifying efficacy and deduces overfitting. Naïve Bayes (NB) provides probabilistic classification with efficiency in large datasets, and K-Nearest Neighbor (KNN) excels in instance-based learning for classification tasks. In DL, CNN are highly efficient at pooling out complex hierarchical patterns from structured data, while Gradient Boosting (GB) and another updated version of GB, Extreme Gradient Boosting (XGB) iteratively improve weak learners for superior predictive performance. These techniques [13,23,24] have demonstrated considerable success in advancing IDS capabilities. However, evolving cyber threats pose ongoing challenges, including scalability, real-time detection, and addressing imbalanced data, which require further research and innovation to enhance future IDS frameworks.

Apart from traditional techniques, bio-inspired frameworks have been quite effective for IDS. Though there have been fewer works in this genre, some have proved the significance of bio-inspired techniques, particularly in anomaly detection. In article [25], a bi-layer structure, with the initial stage eliminates recursive patterns incorporating RF-based methodology (RF-RFE) and optimization approaches that are inspired from biological style, is employed in the following layer. In article [26], the support vector machine was optimized using Whale Optimization (WO), Grey Wolf Optimization

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2025 doi:10.20944/preprints202503.1902.v1

(GWO), and Firefly Optimization (FO) algorithms. However, the detection rates of the irregularities by the optimized methodologies did not demonstrate optimal efficiency. In article [27], a novel combination of Deep Learning and the Dendritic Cell Algorithm (DeepDCA) was integrated into the framework alongside a Neural Network that normalises itself (SNN). However, the implementation utilized a relatively outdated and less diverse dataset.

In addition to traditional methods, FL has become a popular architecture for developing Intrusion Detection Systems (IDS), offering distributes training without allowing access to original data. Various studies have explored different FL-based approaches, including RF-based federated learning [16], shallow Artificial Neural Networks (ANN) with FedAvg for aggregation [17], CNN and RNN [28], and combinations of Multi-Layer Perceptron (MLP) with CNN [29]. Euclides et al. introduced the CICIoT2023 dataset, which has been widely used in experiments involving ML algorithms for classification [11]. Notable studies include [30], which proposed a convoluted structure depending on LSTM for attack detection, and Fray et al., who explored DL models with different stages and functions that activates the model [31]. Maryam et al. deployed unbiased ML algorithms for detecting irregularities and [32], emphasizing fairness and accuracy in model predictions. These efforts highlight the evolution of IDS using FL and advanced machine learning techniques.

3. Methodology

In this section of the article will represent the methodology of the entire research work. This section highlights the research approach, encompassing dataset compilation, data analysis, prior processing, and the proposed framework.

The sections includes the architectural depiction of the proposed framework in the Figure 3. The illustration depicts the process of a federated learning model. It specifies the procedure of initialising a server, disseminating the global framework to local end nodes, changing the model regionally, and transmitting revisions back to the server. The server subsequently incorporates the updates with the Federated Averaging (FedAvg) mechanism. The diagram illustrates the allocation of models and upgrades to features between the server and remote nodes, where multiple users (Client-1, Client-2, Client-3, and Client-4) utilise their datasets to enhance the effectiveness of the models. The remainder of this section will proceed to explore the suggested framework in in greater depth.

3.1. Dataset Compilation

In this work, the CICIoT2023 [11] data set was used for the proposed model. The dataset was created to meet the growing demand for reliable security analytics applications in real-world IoT contexts. The experimental setup comprises of a complete IoT architecture with 105 networked devices that simulate a genuine operational environment. Within this network, 33 different cyberattacks were carried out, each meticulously planned to mirror the changing threat landscape of IoT security. In order to provide diversity and authenticity in attack situations, compromised IoT devices were intentionally used as hostile entities, attacking other IoT devices in the network. The dataset encompasses a wide range of harmful activities, with the goal of serving as a standard for the research community in developing and assessing intrusion detection and mitigation solutions. The assault diversion is shown in the Figure 1.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2025 doi:10.20944/preprints202503.1902.v1

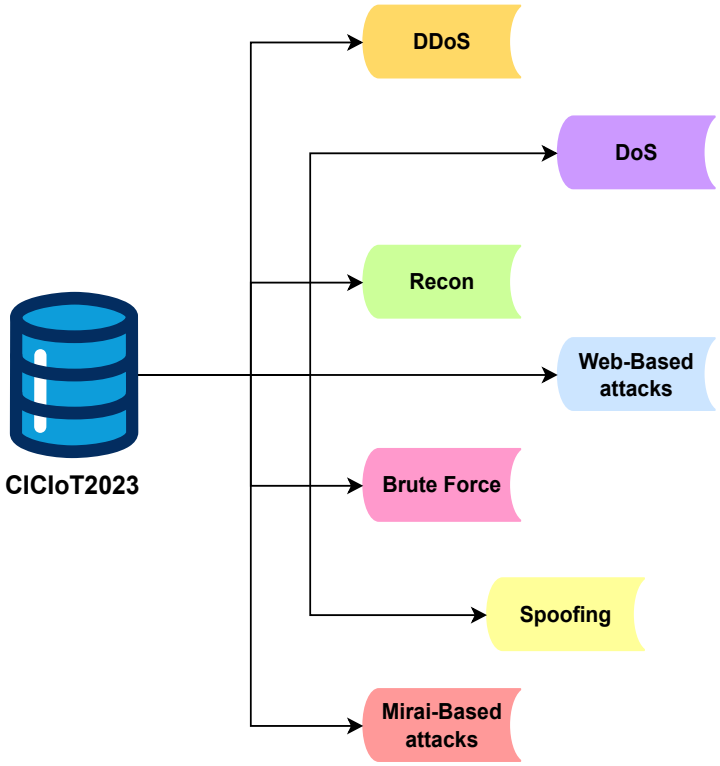


Figure 1. Pictorial representation of the experimented dataset with the attack categories.

The recorded attacks are systematically categorized into seven major classes, each representing a fundamental aspect of IoT security threats:

I. **Distributed Denial of Service (DDoS):** Large-scale flooding assaults meant to harm multiple IoT gadgets simultaneously to exhaust computational resources and disrupt network availability.
II. **Denial of Service (DoS):** Single-source attack tactics designed to overwhelm a specific IoT device, making it unresponsive to valid queries.
III. **Reconnaissance (Recon):** Passive and active network scanning techniques are used to obtain information on vulnerable IoT devices, services, and network setups.
IV. **Web-based Attacks:** Exploitation of IoT web interfaces using security holes in IoT web interfaces, including SQL infiltration, command insertion, and cross- website scripting to gain unauthorized access.
V. **Brute Force Attacks:** Systematic password-guessing attacks targeting IoT authentication mechanisms to compromise credentials and gain illicit control over devices.
VI. **Spoofing Attacks:** Identity forging techniques, such as ARP and IP spoofing, are used to masquerade as legitimate IoT organizations in order to eavesdrop or manipulate communications.
VII. **Mirai-based Attacks:** Malware-driven attacks use the Mirai botnet to exploit vulnerabilities in IoT device security, allowing for large-scale infections and subsequent coordinated cyberattacks.

This dataset is a crucial asset for cybersecurity research, allowing for the creation and testing of advanced machine learning models, intrusion detection systems, and anomaly detection approaches designed specifically for IoT security applications.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2025 doi:10.20944/preprints202503.1902.v1

3.2. *Data Investigation & Pre-Processing*

In this level of the suggested approach, the dataset is thoroughly inspected if the data are in eligible format to use. In order to get an impactful work, the datasets were pre-processed in several steps. The details of the pre-processing is visually represented in the Figure 2 as well as described below,

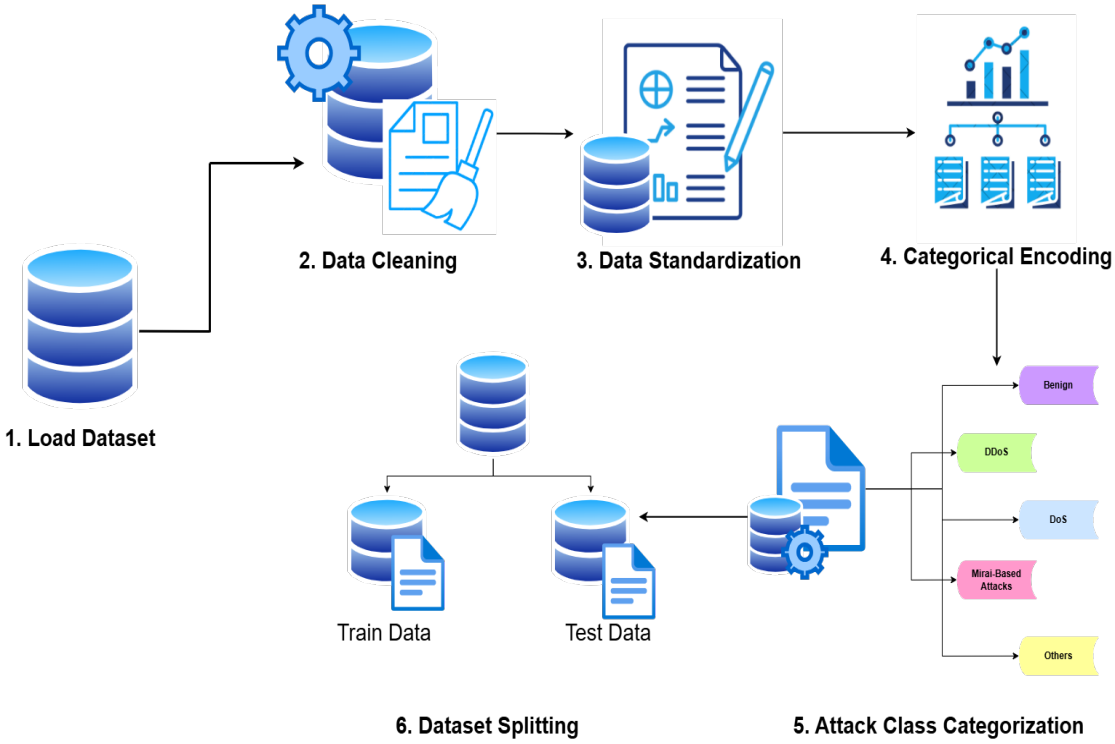


Figure 2. Detailed steps of data preparation for the proposed experiment.

3.2.1. **Data Cleaning**

Proper handling of missing values is important because they could result in errors during model building. Missing rows or columns had been either removed. If the input data contains the same thing multiple times, this can skew the learning outcome. Duplicate values were checked and removed from the dataset.

3.2.2. **Data Type Correcting**

In order to ensure precise calculations, inconsistent data types (e.g., numeric columns erroneously saved as strings) were transformed to their correct formats.

a. **Data Standardizing :** Standardizing the data is a transformation process that increases the integrity and quality of the data that you can use in future calculations. For this study, we used Z-index standardization 1. This method normalizes the data to have zero mean and unit variance.

$$Z = \frac{X - \mu}{\sigma} \tag{1}$$

Here:

- Z denotes the standardised merit,
- X signifies the initial data instance,
- μ represents the mean value of the data,
- σ indicates the standard deviation of the results.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2025 doi:10.20944/preprints202503.1902.v1

b. **Categorical Encoding :** Any machine learning algorithm requires numerical input to perform mathematical operations, so categorical encoding converts categorical data, which contains arbitrary labels or discrete components, into a numerical format. For the encoding technique, we used label encoding in this study. Unlike one hot encoding, label encoding preserves the ordinal nature of categorical variables, allowing for interaction between dummy variables while providing meaningful numeric representation of the synonyms. Treats Categorical Features as a variable to keep the categorical variables interpretability and thus helps in efficient data processing

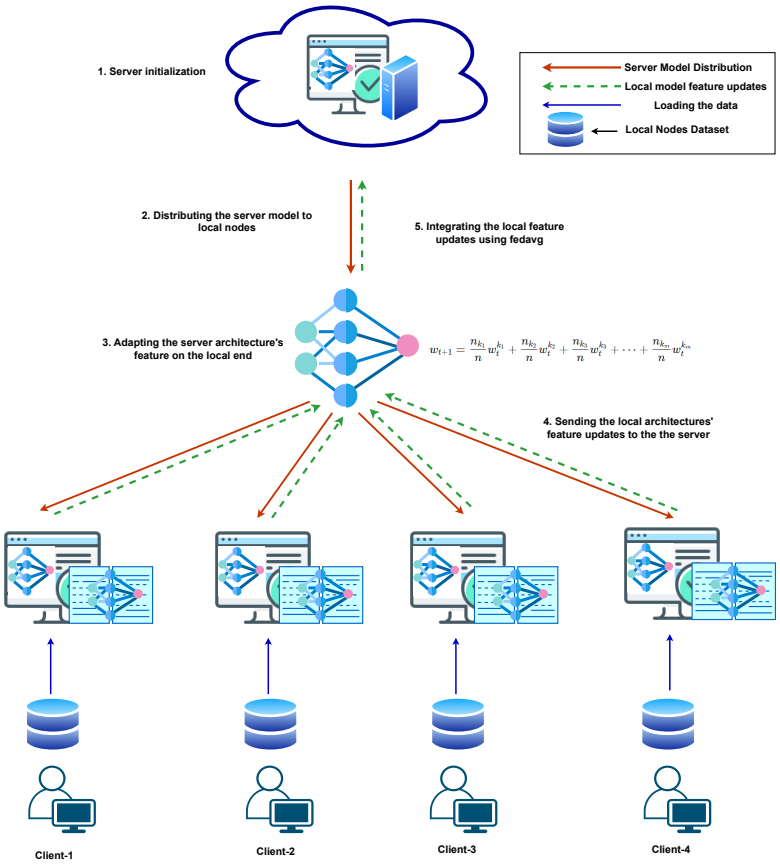


Figure 3. Proposed architectural diagram of FedXAIIDS.

3.2.3. **Class-Conversion**

A robust representation of class conversion is essential in machine learning, as it simplifies the model derived from complex datasets, enhances model performance, and facilitates improved interpretability. The data is simplified and expedited for training due to consolidating classes, which aggregates like assault types. This approach addresses data imbalance by integrating minor attack types into a more significant segment. Moreover, it decreases bias and enhances generalization. Simulataneousy, it aligns with purpose of real-world intrusion detection and increaases the significance of cyber-security. **CICIoT2023** was classified into, **Benign**, **DDoS**, **DoS**, **Mirai**, **Recon** and **Other**.

3.2.4. **Dataset Splitting**

At the last step of pre-processing, the selected data sets are partitioned into two segments. One part is training, consuming 80 percentage of the data and another part is evaluation test, consuming 20 percentage of the data.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2025 doi:10.20944/preprints202503.1902.v1

3.3. Proposed Framework

Our extensive research work experimented with both ML and DL models with two different datasets. We have implemented explainable DL model on CICIoT2023 dataset. The description of the models are given below.

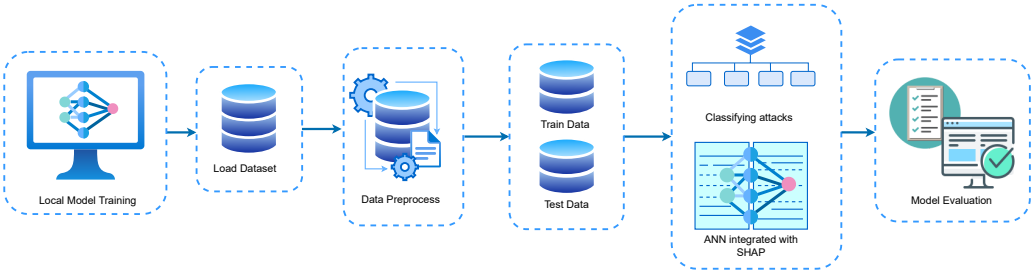


Figure 4. Visual representation of local architecture.

I **FedXAIIDS:** The proposed Federated Learning (FL) model for intrusion detection leverages a distributed architecture, incorporating the devices attached to the local end. These gadgets regionally implement the frameworks on their particular dataset. Later the computing device at the server end integrates these trained schemes. For this study, the CICIoT2023 dataset was distributed across four clients to simulate a federated environment. The following stages define the operation of the proposed model:

(a) **Initialization :** In FL, initialization implies the procedure of establishing the initial universal model prior to the commencement with instruction across various client endpoints. A centralised computer establishes a global architecture and disseminates it to all collaborating peers. Upon acquiring the global framework each client initiates local training utilising their specific data.

(b) **Local Model Training :** Locally model development in FL denotes the procedure whereby every collaborating client (e.g., smartphone or tablet edge nodes) autonomously trains a replica of the global framework on its own dataset prior to transmitting updates to the central server. Every client develops a distinct ANN architecture and explains its results utilising SHAP. ANN is modelled after the architecture of the human brain, utilising layers of interconnected neurones. In this work, the ANN framework is fabricated as following:

i. **Input Layer :** The input layer in an Artificial Neural Network (ANN) with 64 neurons represents the 64 features of the dataset, where each neuron processes a corresponding feature. Using the activation strategy of ReLU 2,

$$f(X) = \max(0, X) \tag{2}$$

, it passes only positive values, ensuring efficient learning and faster convergence. Each feature has calculated weighted inputs, while making negative inputs be 0. Due to its simplicity and non-linearity, ReLU allows for sparse activation and, therefore, scalability, making it suitable for more complicated tasks such as intrusion detection.

ii. **Dropout Layer :** Two dropout layer with a fifty percent reduction in rate helps to reduce over fitting by randomly adjusting fifty percent of the units used for input to zero throughout each training iteration.This technique minimises reliance on single neurones, allowing the network to learn more robust properties.

iii. **Hidden Layer :** The model has two hidden layers to improve learning and feature abstraction. The first hidden layer is made up of 128 neurones with

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2025 doi:10.20944/preprints202503.1902.v1

ReLU activation, which allows the network to record complicated patterns using non-linear transformations. The second hidden layer has 64 neurones and uses ReLU activation to reduce dimensionality while retaining abstraction for better computational efficiency. This layered structure strikes a compromise between learning capacity and processing speed, allowing for deeper pattern identification and more effective generalisation.

iv. **Output Layer :** The layer that is designed to provide outcomes, contains six neuronal cells, each corresponding to one of the six categories of interest in the classifying task. The activation function of softmax is employed, transforming the output into the distribution of probabilities among every category. The softmax function guarantees the sum of probabilities is 1 and enables the model to simply make the most likely class prediction and give more probability to whatever output is more relevant, which makes it popular for multi-class classification problems.

v. **Loss Function :** This framework employs a loss function approach that is know as categorical cross-entropy, quantifying the disparity between the one-hot actual label distribution and the projected distribution of probabilities. This method is used where there is multi-class classification, and it punishes when the prediction is neither close to the labels nor close to the class by solving a loss of negative log likelihood of the actual class. Preventing this loss can help tune the model so it is more left-leaning or right-leaning, which improves its accuracy, resulting in predicted probability aligning better to actual labels.

vi. **Optimizer :** This model employs an optimizing techniques of Adam with a primary learning ratio of 0.001 to take advantage of momentum and adaptation rates for improved training. Adam adapts the learning rates for each individual parameter, which leads to quicker convergence and resilient performance across different types of issues, making it frequently employed in deep learning models.

(c) **Global Aggregator :** The centralised computer consolidates modifications to the model from each client to formulate a unified framework utilising Federated Averaging (FedAvg) 3. This method involves the server systematically calculating a weighted average of the clients' model parameters (weights) according to the magnitude of their local data. FedAvg enhances the global model by integrating varied local insights while maintaining data privacy, establishing it as a fundamental technique in FL [33]. The calculation as follows [33],

$$\theta^{(t+1)} = \sum_{k=1}^{K} \frac{n_k}{N} \theta_k^{(t)} \tag{3}$$

where:

- $\theta^{(t+1)}$ represents the new global model parameters,
- K denotes the number of participating clients,
- n_k is the number of local training samples for client k,
- $N = \sum_{k=1}^{K} n_k$ is the total number of training samples across all selected clients,
- $\theta_k^{(t)}$ represents the locally updated model parameters from client k at round t.

(d) **Explainable AI (XAI) Integration :** Explainability is critical for understanding complex machine learning models. SHAP is a model-agnostic approach that provides local interpretability by assigning importance values to individual features. In this study, the Python library of SHAP was used to calculate SHAP values for a test trial subset. Kernel SHAP, suitable of explaining deep learning models, was employed to inter-

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2025 doi:10.20944/preprints202503.1902.v1

pret predictions and ensure efficient, representative interpretability of the CICIoT2023 dataset [9].

4. Experimental Result

The research experiment was conducted on a portable computer with the configuration features of, Core i5 processor, 4GB main memory, 2.4GHz CPU momentum, and running Microsoft Windows 11. The programming environment utilized Jupyter Notebook and Kaggle Notebook, both with Python 3. Furthermore, the Colab at Google acted as a program framework for conducting the experiment.

4.1. Evaluative Metric

The properties employed for the suggested framework in this investigation include precision, accuracy, recall, alongside F1-score.

ACCURACY : The metric that indicates how frequently an ML approach correctly projects an outcome is known as Accuracy. The division of the number of correct predictions by the total estimation quantity, is implemented to evaluate the accuracy.

$$ACCURACY = \frac{TP^1 + TN^2}{FP^3 + TP + FN^4 + TN} \tag{4}$$

RECALL : The capacity to identify all pertinent instances within the dataset is termed as Recall. It is defined as the ratio of actual positives to the total of true positives and false negatives in the genre of mathematics.

$$RRCALL = \frac{TP}{FN + TP} \tag{5}$$

PRECISION : the metric that represents the efficacy of an ML approach is known as Precision. It reflects the accuracy of the algorithms' positive predictions. It is the ratio of genuine successes to the total number of positive projections.

$$PRECISION = \frac{TP}{FP + TP} \tag{6}$$

F1-SCORE : the balanced average of accuracy and recall is termed as F1-SCORE. Accuracy and Recall are aggregated into a single statistics in order to enhance the comprehension of the efficacy of the suggested framework.

$$F1 - SCORE = \frac{2 \times RECALL \times PRECISION}{PRECISION + RECALL} \tag{7}$$

LOSS : A function called a loss function is a formula of algebra that quantifies the variance between the envisioned results derived from the computerised model and the actual target values. The loss function evaluates the degree to which the forecasts of the model align with the factual information.

$$\text{Loss} = -\sum_{i=1}^{n}\sum_{c=1}^{C} y_{i,c} \log(\hat{y}_{i,c}) \tag{8}$$

4.2. Prior Experiment

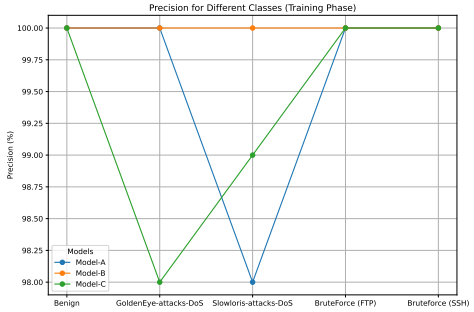
In both Model-A and Model-B, the first layer was meant to extract 20 attributes from the dataset's original 79. This layer was crucial in determining the value of various traits and choosing the most significant ones.

The Deep Component Analysis (DCA) layer in the prior two experimental models refined these features by reducing the 39 retrieved features to ten. This phase improves the ability of the framework to concentrate on the most important elements, hence increasing overall performance and efficiency. The third model, on the other hand, took a completely different strategy. Rather than a two-stage feature extraction, 15 features were recovered directly from the dataset's 79 original attributes. This

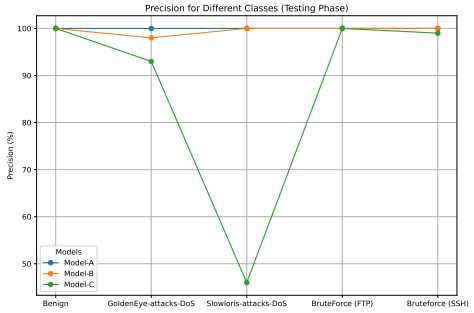
Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2025 doi:10.20944/preprints202503.1902.v1

direct extraction simplifies the model but may lead to differences in feature interpretability and performance compared to the more refined multi-stage process of initial two models.

After the feature extraction process, the models proceed to learning and evaluation using the classifiers discussed in the previous section. It is evident that first framework successfully acquired knowledge of categories of assaults during the learning session, achieving a remarkable learning percentage of 100%. Even the testing phase performance is incredible with a 100% detection rate, showing that the model is extremely accurate and resilient to intrusions. Similar to the previous models, the third framework is extremely efficient with a 100% learning and assessment rate, reflecting its accuracy and reliability in identifying diverse classifications attacks. The images, Figures 5, 6, and 7, depict the results of our previous research employing precision, recall, and F1-score, correspondingly. Here, we can see that although this model has a slightly lower success of 99% in both the learning and assessment elements, it is still admirable. Although marginally less effective than Models A and B, Model-C maintains a high level of accuracy and remains a competitive choice for intrusion detection. The reason behind to switch to our current study is that the previous experiment demonstrated a high probability of overfitting issue of the frameworks as well as the inefficient data quality as the experimented dataset was published quite a long time ago.



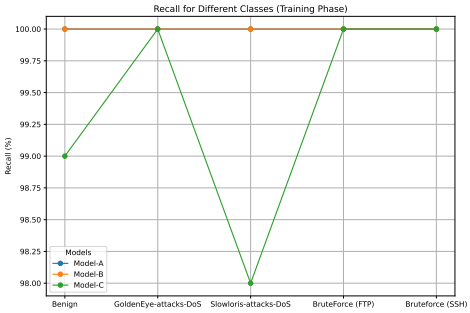
(**a**) Precision graph of Training phase for Experiment-1



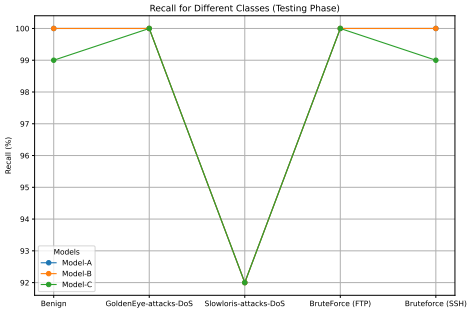
(**b**) Precision graph of Testing phase for Experiment-1

Figure 5. Precision Graph of Experiment-1.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2025 doi:10.20944/preprints202503.1902.v1

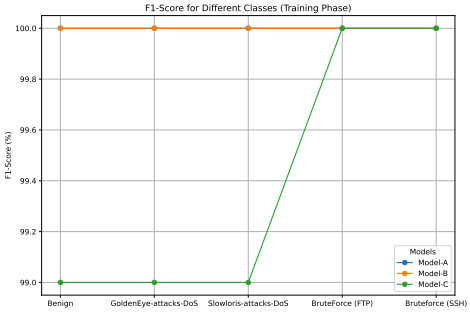


(**a**) Recall graph of Training phase for Experiment-1

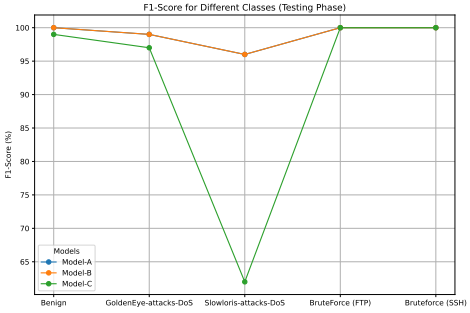


(**b**) Recall graph of Testing phase for Experiment-1

Figure 6. Recall Graph of Experiment-1.



(**a**) F1-Score graph of Training phase for Experiment-1



(**b**) F1-Score graph of Testing phase for Experiment-1

Figure 7. F1-Score Graph of Experiment-1.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2025 doi:10.20944/preprints202503.1902.v1

4.3. FedXAIIDS

The envisioned infrastructure built on FL was allocated among four end-users along with the global framework. The findings have been assembled from the local ends utilizing the aggregator strategy of FedAvg. The resultant products produced during the testing cycles are illustrated in Figure 9. The suggested method has an 88.4% achievement rate during the training session and an 88.2% achievement rate during the testing session. SHAP was employed in order to clarify the significance of the attributes on the anticipated output. Figure 10 illustrates the influence of the characteristics during both the training and testing sessions utilising SHAP. The attributes are represented on the Y- axis ranking from high influence to low influence. The SHAP data are represented on the X-axis . Each dot depicted in the illustration indicate each data of integrated feature. In this context, red signifies a larger importance, while blue denotes the opposite. The illustrated components exert a greater influence on our IDS compared to the other dataset characteristics.

The Figure 8, depicts the accuracy trends of a federated learning model across many clients over a number of repetitions. The horizontal axis depicts the quantity of repetitions, indicating the advancement of training, and the vertical-axis represents accuracy in percentage, representing the model's performance. Every curve in the figure displays an individual client, showcasing how their respective models improve over time. The accuracy deviations could be attributed to differences in data distributions, local model updates, or computing resources. The graph illustrates the converging behavior of federated learning, demonstrating whether all clients attain equal performance levels or if there are differences because of disparities in data or system-related constraints. This approach is important in federated learning because it helps evaluate model consistency across clients, discover potential fairness issues, and ensure that no client suffers a large disadvantage during training. The detected trends can also be used to inform optimization tactics, such as modifying learning rates, enhancing aggregation approaches, and dealing with straggler effects.

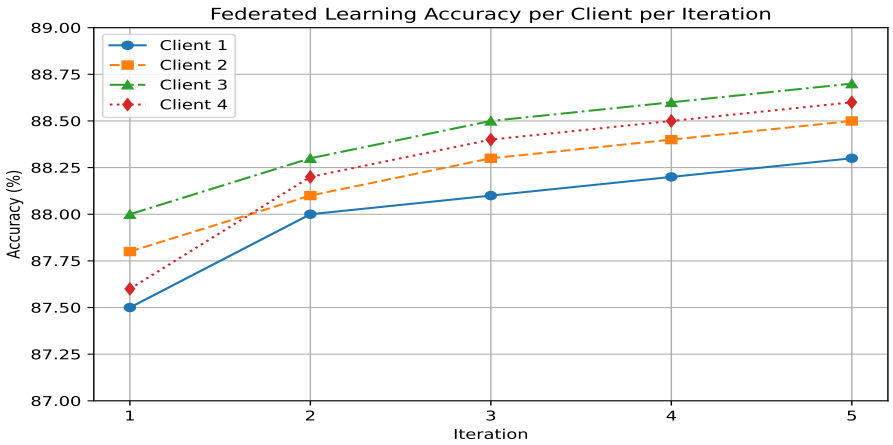
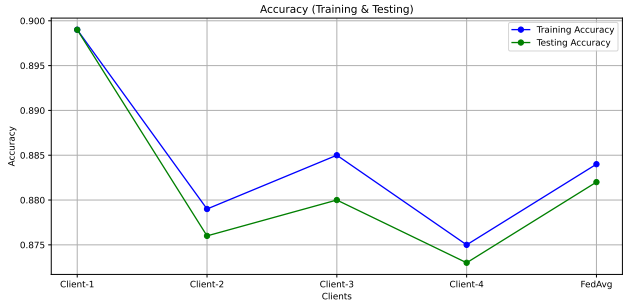


Figure 8. Accuracy trends of federated learning clients over 5 iterations.

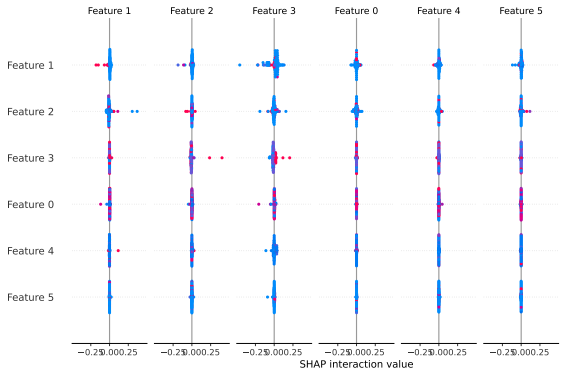
Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2025 doi:10.20944/preprints202503.1902.v1



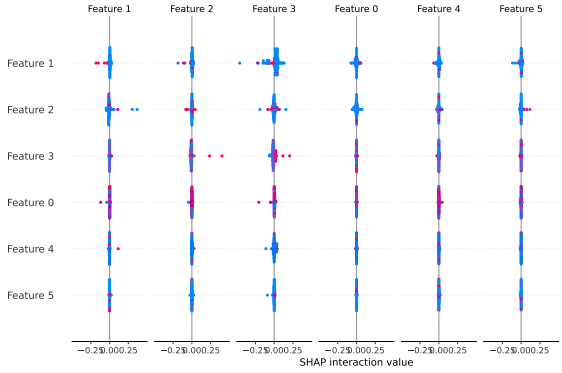
(**a**) Accuracy graph of Testing phase on the clients' end of FedXAIIDS.

(**b**) Loss graph of Testing phase on the clients' end of FedXAI-IDS.

Figure 9. Accuracy Graph of Training and Testing phase of FedXAIIDS.



(**a**) Depiction showing the influence of features utilising SHAP throughout the training session.



(**b**) Depiction showing the influence of features utilising SHAP throughout the testing session.

Figure 10. Graph representation of SHAP interpretation.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2025 doi:10.20944/preprints202503.1902.v1

5. Discussion

The results of this study facilitate multiple promising avenues for the advancement of Intrusion Detection Systems (IDS). Notwithstanding the remarkable outcomes attained with the proposed Federated Learning (FL) framework, the scalability of these models in extensive, heterogeneous contexts continues to offer a significant issue. Future research may concentrate on enhancing federated learning methodologies to accommodate larger datasets, fluctuating network conditions, and immediate threats. This necessitates improving the effectiveness of model aggregation and exploring edge computing solutions to mitigate latency and bandwidth challenges. The existing methodology can be enhanced by investigating hybrid models that integrate various classifiers, perhaps incorporating deep learning frameworks, to augment detection precision, resilience, and adaptation to emerging attack vectors. The incorporation of explainable AI (XAI) methodologies, such as SHAP, has enhanced model transparency; nevertheless, the development of more advanced techniques could yield even more comprehensive and accessible insights, aiding security analysts in interpreting and responding to the system's decisions more effectively. Moreover, although the federated approach provides robust privacy-preserving features, supplementary techniques like differential privacy or secure multi-party computation could be integrated to enhance security without undermining the system's performance. Mitigating the persistent issue of class imbalance in real-world datasets is a vital area for enhancement, with sophisticated data augmentation tactics and semi-supervised learning methodologies offering avenues for improved model training in data-deficient contexts. Furthermore, the real-time implementation of these IDS systems is essential, particularly in dynamic networks, where the capacity to evolve as well as to react to emerging threats instantaneously is crucial. This necessitates the integration of online learning techniques and the development of adaptive models that progress in tandem with evolving assault patterns. Ultimately, the incorporation of these sophisticated IDS methodologies with current cybersecurity frameworks, including SIEM systems and firewalls, could augment their efficacy within a comprehensive security ecosystem, facilitating an integrated detection and response mechanism. Further exploration of these research domains indicates that the integration of federated learning, explainable artificial intelligence, and sophisticated feature refinement techniques offers significant potential for developing robust, scalable, and transparent intrusion detection system solutions that can more effectively safeguard distributed networks against an expanding array of cybersecurity threats.

The findings in the Table 2 show that FedXAIIDS is effective in addressing major intrusion detection concerns, particularly in terms of privacy preservation, explainability, and decentralized learning. While the recommended framework attained training efficacy of 88.4% and testing efficacy of 88.2% on the CICIoT2023 dataset, which is lower than some centralized approaches, its advantages in security, interpretability, and real-world applicability make it a superior choice for modern cybersecurity frameworks. FedXAIIDS, unlike standard IDS models, uses Federated Learning (FL) to ensure that sensitive network data is dispersed across edge devices. This considerably decreases the danger of data breaches and ensures compliance with severe privacy requirements such as GDPR (General Data Protection Regulation). While other high-accuracy models rely on centralized datasets (e.g. Adamova et al. [100% accuracy] and Saadouni et al. [99.83% accuracy]), they compromise data privacy by requiring data aggregation at a central location [34]. Most existing models prioritize accuracy but lack interpretability, making them unsuitable for real-world security operations where decision transparency is crucial. FedXAIIDS integrates SHAP to rank feature importance, allowing cybersecurity professionals to understand attack patterns rather than treating the IDS as a black box. In contrast to high-accuracy systems (e.g. Saadouni et al. and Gulzar et al.) [35,40], which do not explain why specific attacks are categorized, FedXAIIDS delivers reliable insights, allowing for speedier and more informed decision-making. Traditional IDS models perform well on particular datasets, but they frequently suffer in real-world dynamic situations. FedXAIIDS, with its federated learning architecture, learns continually from remote sources, allowing it to adapt to new attack patterns. Centralized models such as Ji et al. (98.27% accuracy) rely on static datasets, making them vulnerable to zero-day attacks and adversarial manipulations [38]. Despite lower accuracy, FedXAIIDS surpasses traditional IDS in

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2025 doi:10.20944/preprints202503.1902.v1

security resilience. It reduces single points of failure, a significant risk in centralized architectures, and integrates smoothly with SIEM systems and firewalls to improve threat detection capabilities. Future advancements, such as distinct confidentiality and safe collective manufacturing, will increase its robustness. While FedXAIIDS does not achieve the highest accuracy, its balance of security, privacy, explainability, and adaptability makes it a more practical and scalable IDS solution for real-world deployment. Future advances in federated model aggregation, machine learning with adaptive features, and hybrid deep learning have the potential to significantly improve its detection performance, driving it to the forefront of next-generation IDSs.

Table 2. Comparison of the evaluation of Experiment-2 with current studies on CICIoT2023.

Ref	Year	Federated Learning applied	Method	Dataset	Performance Metrics (Accuracy)
A. Adamova et al.	2025	Yes	The methodology employs Federated Learning (FL) to enhance IoT security by predicting violations as well as instantaneous evaluation of their cruciality, evaluated on SQL injection and brute force attacks.	CICIOT2023	100% accuracy in predicting SQL injection attacks and 98.25% accuracy for brute force attacks [34].
R. Saadouni et al.	2025	No	It incorporates transfer learning with the beforehand-trained infrastructure of VGG16 for capturing features, along with an optimizer known as Binary Greylag Goose Optimization (BGGO) for feature selection, and a Random Forest classifier for attack detection [35].	CICIOT2023	99.41% accuracy for multiclass classification and 99.83% for binary classification
H. Chen et al.	2025	No	The proposed architecture boosts intrusion recognition in IoT environments utilizing synaptic structures transformation from 1D to 3D. Additionally, imbalance categorization issue is mitigated implement a unique strategy for calculating loss. The experiment was executed on CICIDS2017, and CICIoT2023.	CIC_IDS_2017, CICIOT2023	demonstrated a 88.48% on CICIDS2017 and a 97.69% on CICIoT2023. [36].
J. J. Shirley et al.	2025	No	The proposed methodology integrates an Autoencoder (AE) for feature extraction and dimensionality reduction with a Feedforward Neural Network (FNN) for intrusion classification in IoT networks. A bi-layer balancing scheme boosts identification of minority attacks categories, while the AE-FNN fusion improves accuracy and adaptability to dynamic threats [37].	CICIOT2023	99.55% accuracy in binary classification and 90.91% in multiclass classification.
R. Ji et al.	2025	No	The proposed methodology introduces a hybrid intrusion detection approach for Cyber-Physical Systems (CPSs), integrating AdaBoost and RF atechniques to leverage the advantages of not only bagging but also boosting techniques [38].	CICIOT2023	accurateness of 98.27%, with recall, precision, and F1-score all at 0.98, a false detection rate of 0.0006, along with a testing time of 0.1563 seconds

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2025 doi:10.20944/preprints202503.1902.v1

Sabrina et al.	2025	Yes	The methodology proposes a secure gradients exchange algorithm for distributed intrusion identification in 6G environments, using FL, safeguarded multi-party processing, as well as blockchain to ensure privacy. The model, combining CNN1D and multi-head attention.	CICIOT2023	accuracy of 79.92%, 77.41% identification percentage, and 2.55% of false detection rate [39].
Qawsar et al.	2025	Yes	The methodology introduces a hybrid learning infrastructure for identifying violations in IIoT environments, integrating CNN, LSTM, GRU, and Capsule Networks (CN) [40].	CICIoT 2023 and UNSW_NB15	accuracy of 99.82% on CICIoT 2023 and 95.55% on UNSW_NB15
Damián et al.	2025	No	The methodology presents a Federated Learning-based IDS using a 1D CNN for detecting violations in IoT infrastructures, incorporating privacy techniques like Differential Privacy, Diffie–Hellman Key Exchange, and Homomorphic Encryption [18].	TONIoT, IoT23, BoTIoT, CICIoT2023, CICIoMT2024, RTIoT2022, and Edge-IIoT	The model achieved an estimated accurateness of 97.31%, across the various datasets [].
Ahmad et al.	2024	Yes	This study proposes using Federated Deep Neural Networks (FDNNs) and Explainable AI (XAI) to diagnose and mitigate DDoS assaults in IoT environments, ensuring privacy through federated learning. By integrating XGBoost with SHAP for feature selection [19].	DDoS-ICMP_Flood , DDoS-UDP_Flood , DDoS-TCP_Flood , DDoS-PSHACK_Flood , DDoS-SYN_Flood , DDoS-RSTFINFlood , DDoS-SynonymousIP_Flood , DoS-UDP_Flood , DoS-TCP_Flood , and DoS-SYN_Flood.	the model achieved 99.78% accuracy
JiaMing et al.	2025	Yes	NIDS-FGPA combines federated learning with Paillier encryption for secure training and uses GSA to optimize updates and reduce overhead. A 2D-CNN-BiGRU model handles incomplete data.	Edge-IIoTset and CICIoT2023	Edge-IIoTset and CICIoT2023 datasets exhibit accurateness of 94.5% and 99.2%, correspondingly [41].
FedXAIIDS	2025	Yes	Federated XAI IDS(FedXAIIDS) uses Federated Learning (FL) and SHAP for a privacy-preserving, explainable IDS. An ANN is distributed across four federated clients, aggregated with FedAvg on CICIoT2023.	CICIOT2023	SHAP enhances interpretability, and the model achieved 88.4% training and 88.2% testing accuracy, balancing security, privacy, and trustworthiness.

6. Conclusion

This research introduces Federated XAI IDS, a novel explainable and privacy-preserving intrusion detection system (IDS) that effectively addresses the key limitations of traditional IDS, including high inaccurate positive detection and inaccurate negative detection percentage, lack of interpretability, as well as data privacy concerns. By leveraging Federated Learning (FL) and Shapley Additive Explanations (SHAP), our approach ensures that IDS models can be collaboratively trained across multiple decentralized devices while preserving data privacy by keeping sensitive information on local edge nodes. This decentralized paradigm mitigates security risks associated with centralized approaches, making it a commendatory solution for modern network circumstances. The proposed IDS framework utilizes an Artificial Neural Network (ANN) distributed across four federated clients, with model aggregation performed using FedAvg on the CICIoT2023 dataset. The output highlighted the

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2025 doi:10.20944/preprints202503.1902.v1

efficiency of this approach, achieving 88.4% training accuracy and 88.2% testing accuracy. Additionally, SHAP was incorporated to analyze feature importance, providing a transparent perspective of the most significant attributes influencing model predictions. The ability to rank and interpret feature significance enhances model trustworthiness and supports cybersecurity professionals in making informed decisions. Though our accuracy is lower than other current studies, SHAP analysis ensured the efficiency our result. Our findings demonstrate that the Federated XAI IDS successfully tackles two critical challenges in intrusion detection: explainability and privacy preservation. By integrating federated learning with explainable AI (XAI), this work offers a scalable, interpretable, and secure IDS solution suited for modern cybersecurity applications, particularly in scenarios where sensitive data cannot be centrally shared. Moreover, by promoting more open and responsible intrusion detection framework, the proposed architecture advances the field of reliable AI-driven security solutions. Promising direction for investigation comprise investigating sophisticated federated aggregation tactics, integrating adaptive learning methodologies, and refining model performance in various network contexts. Furthermore, realistic implementation and assessment in extensive, diverse networks will confirm the resilience and applicability of our methodology. We may move closer to a cybersecurity environment that is more safe, considerate of privacy, and interpretable by further improving and developing federated explainable AI-based IDS solutions.

Abbreviations

This document employs the following abbreviations:

IDS	Intrusion Detection System
FL	Federated Learning
XAI	eXplainable AI
AI	Artificial Intelligence
SHAP	SHapley Additive exPLanation
ML	Machine Learning
DL	Deep Learning
ANN	Artificial Nueral Network
FedXAIIDS	Federated Explainable IDS
CNN	Convolutional Neural Network
LSTM	Long Short-Term Memory

References

1. Wang, S.; Asif, M.; Shahzad, M.F.; Ashfaq, M. Data privacy and cybersecurity challenges in the digital transformation of the banking sector. *Computers & security* **2024**, *147*, 104051.
2. Otoum, Y.; Nayak, A. As-ids: Anomaly and signature based ids for the internet of things. *Journal of Network and Systems Management* **2021**, *29*, 23.
3. Okoye, C.C.; Nwankwo, E.E.; Usman, F.O.; Mhlongo, N.Z.; Odeyemi, O.; Ike, C.U. Securing financial data storage: A review of cybersecurity challenges and solutions. *International Journal of Science and Research Archive* **2024**, *11*, 1968–1983.
4. Agrawal, S.; Sarkar, S.; Aouedi, O.; Yenduri, G.; Piamrat, K.; Alazab, M.; Bhattacharya, S.; Maddikunta, P.K.R.; Gadekallu, T.R. Federated learning for intrusion detection system: Concepts, challenges and future directions. *Computer Communications* **2022**, *195*, 346–361.
5. Karatas, G.; Demir, O.; Sahingoz, O.K. Increasing the performance of machine learning-based IDSs on an imbalanced and up-to-date dataset. *IEEE access* **2020**, *8*, 32150–32162.
6. Waswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the NIPS, 2017.
7. Adadi, A.; Berrada, M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access* **2018**, *6*, 52138–52160.
8. Fatema, K.; Anannya, M.; Dey, S.K.; Su, C.; Mazumder, R. Securing Networks: A Deep Learning Approach with Explainable AI (XAI) and Federated Learning for Intrusion Detection. In Proceedings of the International Conference on Data Security and Privacy Protection. Springer, 2024, pp. 260–275.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2025 doi:10.20944/preprints202503.1902.v1

9. Lundberg, S. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874* **2017**.
10. Dong, T.; Li, S.; Qiu, H.; Lu, J. An interpretable federated learning-based network intrusion detection framework. *arXiv preprint arXiv:2201.03134* **2022**.
11. Neto, E.C.P.; Dadkhah, S.; Ferreira, R.; Zohourian, A.; Lu, R.; Ghorbani, A.A. CICIoT2023: A real-time dataset and benchmark for large-scale attacks in IoT environment. *Sensors* **2023**, *23*, 5941.
12. Fatema, K.; Dey, S.K.; Bari, R.; Mazumder, R. A Novel Two-Stage Classification Architecture Integrating Machine Learning and Artificial Immune System for Intrusion Detection on Balanced Dataset. In Proceedings of the International Conference on Information and Communication Technology for Intelligent Systems. Springer, 2024, pp. 179–189.
13. Nguyen, S.N.; Nguyen, V.Q.; Choi, J.; Kim, K. Design and implementation of intrusion detection system using convolutional neural network for DoS detection. In Proceedings of the Proceedings of the 2nd international conference on machine learning and soft computing, 2018, pp. 34–38.
14. de Carvalho Bertoli, G.; Junior, L.A.P.; Saotome, O.; dos Santos, A.L. Generalizing intrusion detection for heterogeneous networks: A stacked-unsupervised federated learning approach. *Computers & Security* **2023**, *127*, 103106.
15. Toldinas, J.; Venčkauskas, A.; Damaševičius, R.; Grigaliūnas, Š.; Morkevičius, N.; Baranauskas, E. A novel approach for network intrusion detection using multistage deep learning image recognition. *Electronics* **2021**, *10*, 1854.
16. Markovic, T.; Leon, M.; Buffoni, D.; Punnekkat, S. Random forest based on federated learning for intrusion detection. In Proceedings of the IFIP international conference on artificial intelligence applications and Innovations. Springer, 2022, pp. 132–144.
17. Lazzarini, R.; Tianfield, H.; Charissis, V. Federated learning for IoT intrusion detection. *Ai* **2023**, *4*, 509–530.
18. Torre, D.; Chennamaneni, A.; Jo, J.; Vyas, G.; Sabrsula, B. Toward Enhancing Privacy Preservation of a Federated Learning CNN Intrusion Detection System in IoT: Method and Empirical Study. *ACM Transactions on Software Engineering and Methodology* **2025**, *34*, 1–48.
19. Almadhor, A.; Altalbe, A.; Bouazzi, I.; Hejaili, A.A.; Kryvinska, N. Strengthening network DDOS attack detection in heterogeneous IoT environment with federated XAI learning approach. *Scientific Reports* **2024**, *14*, 24322.
20. Alsaleh, S.; Menai, M.E.B.; Al-Ahmadi, S. A Heterogeneity-Aware Semi-Decentralized Model for a Lightweight Intrusion Detection System for IoT Networks Based on Federated Learning and BiLSTM. *Sensors* **2025**, *25*, 1039.
21. Bensaid, R.; Labraoui, N.; Ari, A.A.A.; Saidi, H.; Emati, J.H.M.; Maglaras, L. SA-FLIDS: secure and authenticated federated learning-based intelligent network intrusion detection system for smart healthcare. *PeerJ Computer Science* **2024**, *10*, e2414.
22. Sun, S.; Sharma, P.; Nwodo, K.; Stavrou, A.; Wang, H. FedMADE: Robust Federated Learning for Intrusion Detection in IoT Networks Using a Dynamic Aggregation Method. In Proceedings of the International Conference on Information Security. Springer, 2024, pp. 286–306.
23. Al-Imran, M.; Ripon, S.H. Network intrusion detection: an analytical assessment using deep learning and state-of-the-art machine learning models. *International Journal of Computational Intelligence Systems* **2021**, *14*, 200.
24. Arslan, R.S. FastTrafficAnalyzer: An efficient method for intrusion detection systems to analyze network traffic. *Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi* **2021**, *12*, 565–572.
25. Tonni, Z.A.; Mazumder, R. A Novel Feature Selection Technique for Intrusion Detection System Using RF-RFE and Bio-inspired Optimization. In Proceedings of the 2023 57th Annual Conference on Information Sciences and Systems (CISS). IEEE, 2023, pp. 1–6.
26. Haque, N.I.; Khalil, A.A.; Rahman, M.A.; Amini, M.H.; Ahamed, S.I. Biocad: Bio-inspired optimization for classification and anomaly detection in digital healthcare systems. In Proceedings of the 2021 IEEE International Conference on Digital Health (ICDH). IEEE, 2021, pp. 48–58.
27. Aldhaheri, S.; Alghazzawi, D.; Cheng, L.; Alzahrani, B.; Al-Barakati, A. DeepDCA: novel network-based detection of IoT attacks using artificial immune system. *Applied Sciences* **2020**, *10*, 1909.
28. Rashid, M.M.; Khan, S.U.; Eusufzai, F.; Redwan, M.A.; Sabuj, S.R.; Elsharief, M. A federated learning-based approach for improving intrusion detection in industrial internet of things networks. *Network* **2023**, *3*, 158–179.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2025 doi:10.20944/preprints202503.1902.v1

29. Liu, W.; Xu, X.; Wu, L.; Qi, L.; Jolfaei, A.; Ding, W.; Khosravi, M.R. Intrusion detection for maritime transportation systems with batch federated aggregation. *IEEE Transactions on Intelligent Transportation Systems* **2022**, *24*, 2503–2514.
30. Yaras, S.; Dener, M. IoT-Based Intrusion Detection System Using New Hybrid Deep Learning Algorithm. *Electronics* **2024**, *13*, 1053.
31. Becerra-Suarez, F.L.; Tuesta-Monteza, V.A.; Mejia-Cabrera, H.I.; Arcila-Diaz, J. Performance Evaluation of Deep Learning Models for Classifying Cybersecurity Attacks in IoT Networks. In Proceedings of the Informatics. MDPI, 2024, Vol. 11, p. 32.
32. Khan, M.M.; Alkhathami, M. Anomaly detection in IoT-based healthcare: machine learning for enhanced security. *Scientific Reports* **2024**, *14*, 5872.
33. McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; y Arcas, B.A. Communication-efficient learning of deep networks from decentralized data. In Proceedings of the Artificial intelligence and statistics. PMLR, 2017, pp. 1273–1282.
34. Adamova, A.; Zhukabayeva, T.; Mukanova, Z.; Oralbekova, Z. Enhancing internet of things security against structured query language injection and brute force attacks through federated learning. *International Journal of Electrical & Computer Engineering (2088-8708)* **2025**, *15*.
35. Saadouni, R.; Gherbi, C.; Aliouat, Z.; Harbi, Y.; Khacha, A.; Mabed, H. Securing smart agriculture networks using bio-inspired feature selection and transfer learning for effective image-based intrusion detection. *Internet of Things* **2025**, *29*, 101422.
36. Chen, H.; Wang, Z.; Yang, S.; Luo, X.; He, D.; Chan, S. Intrusion detection using synaptic intelligent convolutional neural networks for dynamic Internet of Things environments. *Alexandria Engineering Journal* **2025**, *111*, 78–91.
37. Shirley, J.J.; Priya, M. An Adaptive Intrusion Detection System for Evolving IoT Threats: An Autoencoder-FNN Fusion. *IEEE Access* **2025**.
38. Ji, R.; Selwal, A.; Kumar, N.; Padha, D. Cascading Bagging and Boosting Ensemble Methods for Intrusion Detection in Cyber-Physical Systems. *Security and Privacy* **2025**, *8*, e497.
39. Sakraoui, S.; Ahmim, A.; Derdour, M.; Ahmim, M.; Namane, S.; Dhaou, I.B. FBMP-IDS: FL-based blockchain-powered lightweight MPC-secured IDS for 6G networks. *IEEE Access* **2024**.
40. Gulzar, Q.; Mustafa, K. Enhancing network security in industrial IoT environments: a DeepCLG hybrid learning model for cyberattack detection. *International Journal of Machine Learning and Cybernetics* **2025**, pp. 1–19.
41. Wang, J.; Yang, K.; Li, M. NIDS-FGPA: A federated learning network intrusion detection algorithm based on secure aggregation of gradient similarity models. *PloS one* **2024**, *19*, e0308639.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.