Emerging Trends in Engineering and Sustainability (ETES) Volume 01 (No. 01), 2025, pp. 37-44

Emerging Trends in Engineering and Sustainability (ETES)

https://etes.alnaji-uni.edu.iq

https://doi.org/10.65505/ETES-25-0011

ISSN 3105-1723 (Print)
ISSN 3105-1731 (Online)

A Comparative Analysis of Federated Learning on Non-IID Data for an Intrusion Detection System

Rusul Tareq Khudhair [1*], Chin Hock Goh [2], Asmidar Bte Abu-Bakar [3]

[1] College of Computing & Informatics (CCI) Universiti Tenaga Nasional, Kajang, Malaysia.

Article Info

Received: 08 November 2025
Revised: 27 December 2025
Accepted: 28 December 2025
Available online:
31 December. 2025

Keywords:

Federated learning;
FedAvg algorithm;
Scaffold algorithm;
CICIDS2017 datasets;
Intrusion Detection System (IDS).

ABSTRACT

Federated learning (FL) offers a robust and privacy-preserving approach for developing collaborative intrusion detection systems (IDS). However, statistical variance severely hinders its practical application. Although privacy-preserving federated learning models have been used to develop intrusion detection systems for cyberattacks, problems arise when statistical variance is present. In practice, the performance of the FedAvg algorithm is significantly affected by the heterogeneous distribution of customer data in a real-world network. This distribution causes skewness among customer data, resulting in poor detection accuracy, delayed convergence, and model instability. In this paper, we conducted a comprehensive comparison of the Scaffold algorithm with the FedAvg baseline using the CICIDS2017 datasets. Because the Scaffold algorithm addresses the client skew problem using control variables, it is considered a state-of-the-art federated optimization technique under the heterogeneous partitioning approach. This paper documents the importance of using the Scaffold algorithm as a reliable and essential tool for building high-performance detection systems in a variety of scientific settings. Therefore, our results demonstrate that Scaffold achieved more stable convergence and outperformed FedAvg, with a 15.1% increase in F1-score and a 13.6% higher overall accuracy under highly skewed data distributions. The present evaluation process operates through simulation testing, but physical testbed implementation remains essential for future work to evaluate real-world deployment challenges.

1. Introduction

The continuously growing deployment of Internet of Things (IoT) devices and virtualized enterprise networks has opened a large attack space for advanced cyber threats.[1][2]. Another important aspect for the protection of these infrastructures is Intrusion Detection Systems (IDS) [3]. However, classical centralized IDS architectures are not well designed. They necessitate the collection of a huge volume of raw network traffic onto a centralized server, leading to severe privacy threats, communication bottlenecks and a single point of failure.

Federated Learning (FL) has recently emerged as a promising paradigm to address some of these challenges. By learning models jointly among decentralized clients without exchanging raw data, FL provides a strong privacy-preserving approach [4].

Therefore, FL is being extensively studied for applying distributed IDS. But a key, and often neglected, issue impeding the practical application of FL is statistical heterogeneity. In practice, a data distribution over the clients is rarely IID. Data is intrinsically non-IID, as distinct clients (e.g., separate companies' branches or IoT networks) encounter different types and amounts of attacks

that cause a collection of extremely biased local datasets[5]. Learn a good global model for non-IID data when training with standard FL algorithms (e.g., FedAvg), and that this can in part be attributed to what presents refer to as client-drift. Local models disagree significantly, causing unstable convergence, low global model accuracy and inability of intrusion detection. To combat this instability, more advanced FL algorithms have been developed. Scaffold is a modern approach that corrects for the client-drift by using control variates to handle the discrepancy between local and global model updates.

Theoretically appealing, the practical effect of the scaffold in high-stakes (non-) IID cases (e.g., for ID) is not fully studied. This work presents a machine learning-based Intrusion Detection System (IDS), abbreviated as FD-IDS, which achieves intrusion detection in IoT environments while maintaining data privacy. Artificial intelligence is used to select features, improve their quality, reduce redundancy in high-dimensional data, and enhance model training efficiency and detection performance[6]. This paper addresses this gap with detailed experimental evidence. This research presents an evaluation of the FedAvg basic algorithm versus the advanced Scaffold algorithm for a unified intrusion detection system, using training on the CICIDS2017 dataset under realistic scenarios that do not rely on intruder identification.

The contributions of this paper are as follows: Implementation, deployment, and validation Achieving a high level of detection. Providing a comparative analysis and laying a solid foundation for further research in the field of practical, decentralized, and reliable intrusion detection. The structure of this paper is as follows: Section 2 provides a review of the related work on IDS. Section 3 details the proposed methodology. selecting the datasets, pre-processing, and FL algorithms. Section 4 Experimental results are provided and discussed. Section 5, this research concludes and discusses. Finally, future work in Section 6.

2. Related Work

Federated Learning applied to Intrusion Detection Systems (IDS) is an area of research growing at a fast pace and is motivated by the necessity for security solutions that are privacy-preserving in distributed networks. The pure centralized IDS models are intolerable, and researchers have proved that FL provides an approach to collaborate in the joint training of a high-accuracy IDS without revealing private network traffic data. However, the most challenging issue preventing these systems from being widely used in practice is statistical heterogeneity (or non-IID data).

This phenomenon has been surveyed in a few papers that point out the fact that distributions on real-world networks are inherently skewed. Indeed, for instance, various branches of an organization or sub-networks (in our case, IoT) will suffer from different types and volumes of attacks. Therefore, in practice, expect the absence of independent and identical distributions to be very noticeable. This lack of independent and identical data leads to poor performance of standard fuzzy logic algorithms (such as FedAvg [5]). Furthermore, client skewness results in training instability and significantly reduced detection accuracy [7].

Some researchers have started to suggest a remedy to this trouble. For example, Peng et al.[5] suggested applying Knowledge Distillation for a better stability of training in non-IID IoT setups. Other studies have investigated variations in the aggregation strategy or model architecture. Although these works confirm the severity of the non-IID problem, a crucial missing piece remains: no direct, empirical comparison of state-of-the-art drift-correcting algorithms has been performed with respect to the standard FedAvg baseline on common cybersecurity tasks. Various state-of-the-art algorithms have been developed to deviate from this IID setting, e.g., Scaffold [3], which uses control variates to adjust for client drift. It is, however, not yet clear what its empirical performances are in practice for intrusion detection[8].

In a recent publication, [9] tested FL algorithms for identifying IoT threats using the CICIoT2023 dataset. Their comparison study revealed that while Scaffold was stable across various data distributions, FedProx was more adept at adjusting to non-IID situations than FedAvg. The scientists did note that Scaffold still must be adjusted to boost its effectiveness and that solving scalability problems in large-scale networks is still challenging. This finding

motivates our current efforts to extensively analyse Scaffold's performance and establish a solid baseline against client drift.

Some modern methods have been developed to circumvent this identical and independent arrangement assumption. For example, the Scaffold method [3] was constructed by slightly modifying the control variables to identify the penetration. In fact, the research work scheduled for 2025 [Ref_Xu] is addressing one of the significant challenges, which is heterogeneity in machine learning contexts, by providing a review of the latest approaches for mitigating the consumer skewness at the sample level. Nevertheless, it is still controversial to compare the experimental penetration detection systems' performance in specific scenarios[10].

Therefore, the contribution of this paper is a direct comparison between FedAvg [5] and Scaffold [3]. aiming to isolate the impact of client drift, and provided empirical evidence of Scaffold's strong leadership, setting a firm baseline for all further work in practical federated IDS.

Table 1. Summary of literature

Reference	Technique	Dataset	Focus Area	Key Findings & Limitations
Peng et al.	FL + Knowledge Distillation	IoT Networks (Non-IID)	efficiency, training stability	proposed using Knowledge Distillation (KD) to improve stability in heterogeneous IoT environments, but he focused more on model architecture than direct deviation correction algorithms.
Scaffold	Scaffold Control Variates	Generic FL Tasks	Client Drift	A theoretical solution for skew correction using control variables has been suggested, but its experimental effectiveness in penetration detection scenarios is still unclear.
Ref (e.g., Gad)	FedAvg, FedProx, Scaffold	CICIoT2023	IoT attacks, statistician not present, Consumer drift	FeedProx showed good performance with non-independent, matched-distributed (non-IID) data. The scaffold was stable but needed fine-tuning, and it faced challenges with scalability.
Ref_Xu (2025)	Review of Mitigation Methods	General ML Contexts	Consumer Skewness	A recent review of sample-level skew mitigation methods confirms that dealing with heterogeneity is still a major challenge.
Proposed Method	Scaffold vs. FedAvg	Common Cybersecurity Tasks (CICIDS2017)	isolating the impact of customer drift	This study offers a direct experimental comparison, which is currently missing in published research, to establish a solid basis for Scaffold's performance in penetration detection systems under non-independent, matched-distributed data conditions.

3. Methodology

Based on real-world network scenarios, a simulation framework is used to empirically assess how well federated learning algorithms function under statistical variance. The methodology includes dataset preparation and selection, comparative federated learning algorithms, and a non-inherent data partitioning (non-IID) strategy.

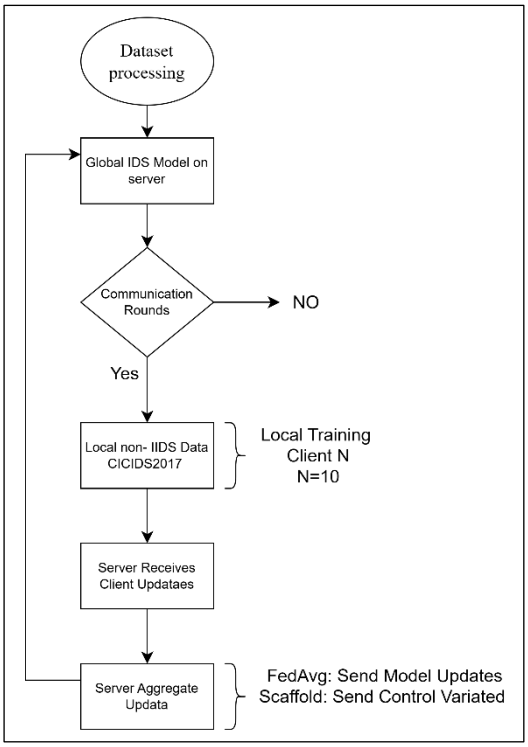

Figure 1. Flowchart of the Federated learning Intrusion Detection System

3.1 Dataset Description

A publicly available dataset was used for network penetration testing, CICIDS2017: A comprehensive dataset with balanced traffic, encompassing a wide range of contemporary attack scenarios and balanced, healthy traffic[11].

To purify the data, first excluded samples with unlimited or missing values. implemented used one-hot coding to convert categorical attributes (such as protocol and service) into numerical representations. To prevent any single attribute from dominating during training, all numerical properties were standardized within a common range [0, 1] using a max-min divider. Finally, created a binary (healthy vs. attack) classification for the penetration testing model by combining the different attack classifications. [12]

3.2. Federated Configuration and Model Architecture

Our global model, which can be deployed on decentralized clients, was a lightweight Multi-Layer Perceptron (MLP). Two hidden layers with ReLU activation functions and an output layer with SoftMax activation make up the model[13].

The CLIENTS option in our code was used to mimic a federated system with a central server and N=10 clients. By dividing the training data among clients using a pathological non-IID technique based on a Dirichlet distribution (Dir (α)), the fundamental problem of statistical heterogeneity was simulated. [14]

Create a highly realistic and hard environment for the FL algorithms by setting the concentration parameter alpha to a low value (e.g., alpha=0.5), which guarantees that just a few distinct classes (attack types) dominate each client's local dataset. [15][16]

3.3. Implemented Algorithms

The given code explains the direct opposition of our proposed methodology to a sophisticated solution (Scaffold) to the traditional baseline (FedAvg) [7].

FedAvg, which is also known as the baseline algorithm, is used as a benchmark for comparative analysis. In this paradigm, the server sums up current local models that are sent by the participating clients and then calculates a weighted average—depending on the sample size of each client—to combine the resulting global model. Nonetheless, this seemingly simple process is highly vulnerable to client drift when heterogeneity of data (non-IID) is present, triggering instability and a subsequent decline in predictive accuracy [8].

Scaffold (Controlled Average): This is an advanced algorithm that has been applied to reduce client drift. Scaffold is a complex mechanism through which it achieves its goals, which is known as control variables. These variables act as error correctors and estimate and track the level of drift to the training direction at client and server levels. Through correction of this divergence at every iteration, Scaffold ensures that all the agents all come to the same, common global goal, which generates faster, more stable, and more precise convergence. [7]

3.4. Evaluation Metrics

Our evaluation function calculates standard classification metrics, essential for imbalanced IDS datasets, to accurately evaluate the performance of the two algorithms:

- Accuracy: The proportion of cases accurately classified. [17]

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

- Precision: A measure of the quality of prediction, calculated as the ratio of true positives to all predicted positives[18].

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

- Recall: The completeness of detection is measured by the ratio of true positives to all true positives [19].

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

- F1 score: A reliable statistic for imbalanced data, calculated as a harmonic mean of precision and recall[20].

$$\text{F1 score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

4. Results and Discussion

Based on the results from our experimental framework, we provide and evaluate the FedAvg and Scaffold performance comparison in this section.

4.1 Experimental Setup

The simulation was conducted on a high-performance workstation equipped with a 13th Gen Intel Core i7-13620H processor (10 cores, up to 4.90 GHz) and 16 GB of RAM. The experiments were implemented using Python within a controlled environment to ensure reproducibility.

4.2 Confusion Matrix for Algorithms

4.2.1 Confusion Matrix for FedAvg(non-IID)

To provide a deeper analysis of error types, Figure 2 shows the confusion matrix for FedAvg.

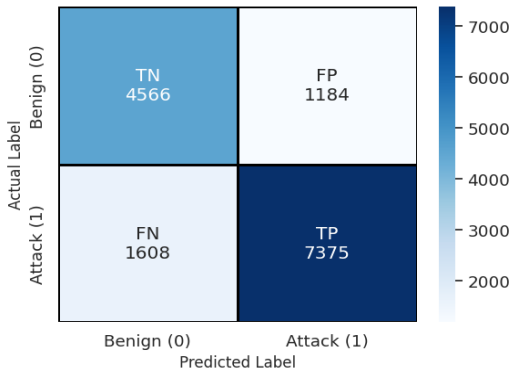

Figure 2. Confusion Matrix for FedAvg (Non-IID)

The confusion matrix for FedAvg (Non-IID) in Figure 2 above showed many false positives (FP) (1184) and, more importantly, 1608 false negatives (FN), meaning it missed more than 1600 real attacks.

4.2.2 Confusion Matrix for Scaffold (non-IID)

The confusion matrix of the Scaffold algorithm has been implemented and will also be used to perform a more in-depth analysis of error types. Figure 3 below shows the confusion matrix of the Scaffold algorithm.

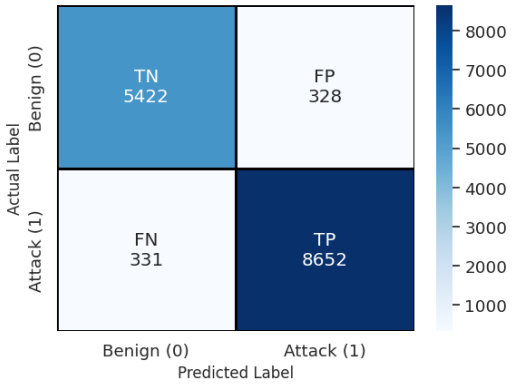

Figure 3. Confusion Matrix for Scaffold (Non-IID)

The structure (Figure 3) above has greatly reduced both types of error, especially serious false negative results (331), confirming that it is a more reliable model.

4.2 Performance Comparison on Non-IID Data

The CICIDS2017 dataset under a non-IID setting. To simulate statistical heterogeneity among N=10 clients, data partition was generated using a Dirichlet distribution (α=0.5). The experiments were conducted over T=100 communication rounds. For local training, E=5 epochs per round were utilized, employing the Adaptive Moment Estimation (Adam) optimizer with a learning rate of lr=0.01. Our analysis begins with the overall convergence behavior; this visual evidence corroborates the final performance metrics summarized in Figure 4 below.

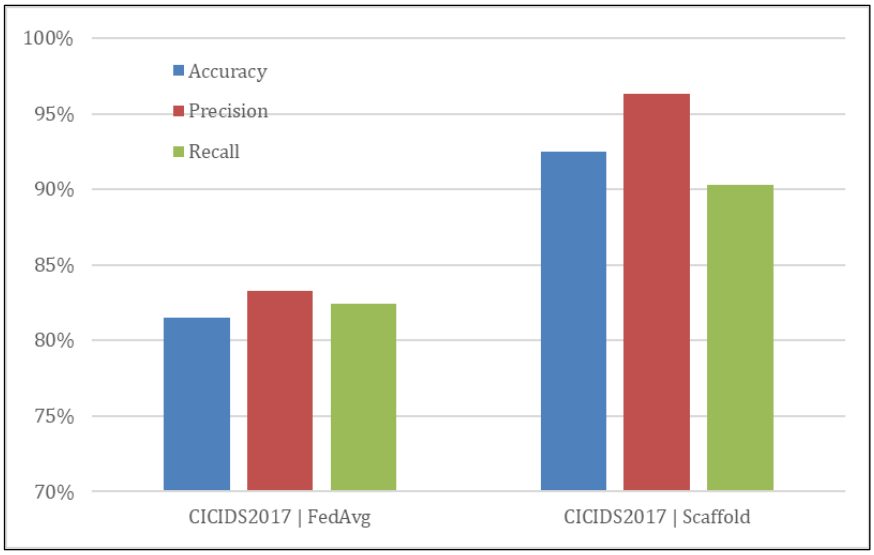

Figure 4. Comparison Model Performance on Non-IID Dataset

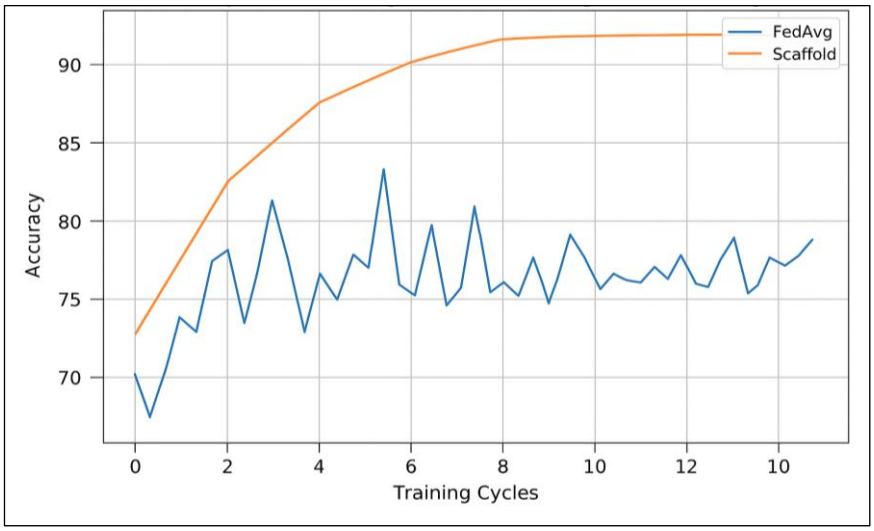

Figure 5. Comparison of the FedAvg and Scaffold algorithms

Figure 5 shows a significant difference in the performance of the FedAvg algorithm (indicated by the blue line) and the Scaffold algorithm (indicated by the orange line). The plotted performance is highly volatile in nature, with sudden changes in the accuracy of classification. The model therefore ends up converging to a relatively low plateau in accuracy, in the range of 81 to 82 percent, because of its inefficient convergence behavior. The relative performance of the FedAvg and Scaffold algorithms is shown below. Two salient conclusions can be drawn out of these observations. To begin with, the FedAvg method does not provide a reliable intrusion detection model in the real-life heterogeneous data environment. Its convergence is very unstable, as shown in Figure 5, demonstrating the adverse effect of client skewness on the aggregate model performance. Second, the Scaffold approach lowers down this bias, achieving significantly good results on the F1 metric (e.g., 93.19% versus 82.85% on the CICIDS2017 benchmark) and yielding better overall accuracy. The scaffold algorithm (orange line) shows a consistent and predictable convergence path with high accuracy (over 90 percent), achieving it in a short time and being consistent across training epochs with low variance.

5. Conclusions

The paper explores a fundamental issue with the standardized intrusion detection systems (IDS) in relation to statistical heterogeneity (non-IID). Experimental evidence demonstrates that the

FedAvg algorithm is poorly applied in heterogeneous real-life situations because it does not converge to a stable and accurate model. Client-side data skew impairs the performance of the algorithm. The Scaffold model uses control variables to mitigate client skewness and account for local variations, resulting in better, more stable, and more convergent performance. Therefore, Scaffold outperformed FedAvg, achieving 13.6% higher accuracy and 15.1% higher F1 scores under highly skewed data distributions, according to our experimental results on the CICIDS2017 datasets. As a reliable, efficient, and essential foundation for any realistic and high-performance standardized intrusion detection system, this study validates the Scaffold model. Currently, the work will be carried out through simulation, and in the future, it will focus on deploying this framework on real-world Internet of Things testing platforms to assess physical constraints.

Future research will focus on developing this robust explainable AI (FL) model, built on the Scaffold platform, to create a new and reliable framework for explainable AI (XAI).

Acknowledgements

The authors thank the lab team at University Tenaga Nasional for their assist in conducting the experiments of the current research.

Conflict of interest

The authors declare no conflicts of interest concerning this research.

Funding

No funding was received for conducting this study.

Author Contribution

R.T. Khudhair and C.H. Goh proposed the research problem.

R.T. Khudhair and A.B. Abu-Bakar developed the theoretical framework.

R.T. Khudhair, C.H. Goh, and A.B. Abu-Bakar verified the analytical methods.

All authors participated in the discussion of the results and contributed to writing the manuscript.

AI Declaration Statement

The authors confirm that the manuscript has been written without the assistance of generative AI or AI-based writing tools.

References

[1] Xu, Binbin, et al. “Cross-Component Transferable Transformer Pipeline Obeying Dynamic Seesaw for Rotating Machinery with Imbalanced Data.” *Sensors*, vol. 23, no. 17, 25 Aug. 2023, pp. 7431–7431, https://doi.org/10.3390/s23177431.

[2] Busra Buyuktanir, et al. “Federated Learning in Intrusion Detection: Advancements, Applications, and Future Directions.” *Cluster Computing*, vol. 28, no. 7, 4 Aug. 2025, https://doi.org/10.1007/s10586-025-05325-w.

[3] R. Lazzarini, H. Tianfield, and V. Charissis, “Federated Learning for IoT Intrusion Detection,” *AI*, vol. 4, no. 3, pp. 509–530, Sep. 2023, https://doi.org/10.3390/ai4030028

[4] Li, Tian, et al. “Federated Learning: Challenges, Methods, and Future Directions.” *IEEE Signal Processing Magazine*, vol. 37, no. 3, May 2020, pp. 50–60.https://doi.org/10.1109/msp.2020.2975749.

[5] Zeng, Yan, et al. “Adaptive Federated Learning with Non-IID Data.” *The Computer Journal*, vol. 10.1093/comjnl/bxac118, no. 11, November 2023, Pages 2758–2772, 30 Sept. 2022, https://doi.org/10.1093/comjnl/bxac118.

[6] Peng, Haonan, et al. “FD-IDS: Federated Learning with Knowledge Distillation for Intrusion Detection in Non-IID IoT Environments.” *Sensors*, vol. 25, no. 14, 10 July 2025, p. 4309, https://doi.org/10.3390/s25144309.

[7] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, “Communication-Efficient Learning of Deep Networks from Decentralized Data,” proceedings.mlr.press, Apr. 10, 2017. https://proceedings.mlr.press/v54/mcmahan17a?ref=https://githubhelp.com

[8] Karimireddy, Sai Praneeth, et al.

"SCAFFOLD: Stochastic Controlled Averaging for Federated Learning." *ArXiv:1910.06378 [Cs, Math, Stat]*, vol. 1910.06378, no. Sai Praneeth Karimireddy, 9 Apr. 2021, arxiv.org/abs/1910.06378.

[9] E. Gad, Zubair Md Fadlullah, and M. M. Fouda, "A Robust Federated Learning Approach for Combating Attacks Against IoT Systems Under Non-IID Challenges," arXiv (Cornell University), pp. 1–6, May 2024, https://doi.org/10.1109/smartnets61466.2024.10577749.

[10] Xu, Haoran, et al. "Federated Learning with Sample-Level Client Drift Mitigation." *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 20, 11 Apr. 2025, pp. 21752–21760, ojs.aaai.org/index.php/AAAI/article/view/35480, https://doi.org/10.1609/aaai.v39i20.35480

[11] Jose, Jinsi, and Deepa V. Jose. "Deep Learning Algorithms for Intrusion Detection Systems in Internet of Things Using CIC-IDS 2017 Dataset." *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, no. 1, 1 Feb. 2023, p. 1134, https://doi.org/10.11591/ijece.v13i1.pp1134-1141.

[12] Kurniabudi, et al. "CICIDS-2017 Dataset Feature Analysis with Information Gain for Anomaly Detection." *IEEE Access*, vol. 8, no. 10.1016/j.ins.2021.06.039, 2020, pp. 132911–132921, ieeexplore.ieee.org/document/9142219, https://doi.org/10.1109/ACCESS.2020.3009843.

[13] Antonio Pietrabissa, and Danilo Menegatti. *Decentralised Learning for Intelligent Control Systems*. 2 Jan. 2022.

[14] S. Khan et al., "Bilevel Hyperparameter Optimization and Neural Architecture Search for Enhanced Breast Cancer Detection in Smart Hospitals Interconnected with Decentralized Federated Learning Environment," IEEE Access, vol. 12, pp. 63618–63628, 2024, doi: https://doi.org/10.1109/access.2024.3392572.

[15] Kazi Fatema, et al. "Federated XAI IDS: An Explainable and Safeguarding Privacy Approach to Detect Intrusion Combining Federated Learning and SHAP." *Future Internet*, vol. 17, no. 6, 26 May 2025, pp. 234–234, www.mdpi.com/1999-5903/17/6/234, https://doi.org/10.3390/fi17060234.

[16] T.-M. H. Hsu, H. Qi, and M. Brown, "Measuring the Effects of Non-Identical Data Distribution for Federated Visual Classification," *arXiv:1909.06335 [cs, stat]*, Sep. 2019, Available: https://arxiv.org/abs/1909.06335.

[17] Swets, J. "Measuring the Accuracy of Diagnostic Systems." *Science*, vol. 240, no. 4857, 3 June 1988, pp. 1285–1293, https://doi.org/10.1126/science.3287615.

[18] Ashish Mohare. "Accuracy." *Scribd*, 2025, www.scribd.com/document/428946227/Accuracy. Accessed 26 Dec. 2025.

[19] Fedorchenko, Elena, et al. "Comparative Review of the Intrusion Detection Systems Based on Federated Learning: Advantages and Open Challenges." *Algorithms*, vol. 15, no. 7, 15 July 2022, p. 247, https://doi.org/10.3390/a15070247.

[20] Humphrey, A, et al. "Machine-Learning Classification of Astronomical Sources: Estimating F1-Score in the Absence of Ground Truth." *Royal Astronomical Society*, vol. 517, no. 1, 16 Oct. 2022, pp. L116–L120, https://doi.org/10.1093/mnrasl/slac120.