IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Performance Analysis of Explainable Federated Learning for Privacy-Preserving Intrusion Detection in IoT Networks

Dr. Pankaj Saxena
Professor
R.B.S. Management Technical Campus, Agra (India)

Abstract—Collaborative learning has its benefits for intrusion detection in IoT, but data has to be sent to one location where attack data could expose operational data. The primary goal is to design an explainable, federated learning integrated Intrusion Detection System (IDS) capable of attack detection without collaborative learning, while keeping data on clients' premises. The Edge-IIoTset dataset that is prepared for machine learning and ready for reproduction was used for this study (157,800 rows, 63 columns). It contains 814 exact duplicates and an analytical sample of 60,000 normal operational data and 14 attack instances. Among the 34 features, leakage was reduced. There were 5 clients for FedAvg and IID as well as non-IID. The Dirichlet distribution was set at $\alpha = 0.3$. The evaluation was performed against five instances of central learning. The results for central learning were the highest for AUC (0.967 ± 0.002), MCC (0.624 ± 0.012), and F1 (0.906 ± 0.009). For IID-FL, AUC was 0.938 ± 0.003, F1 was 0.870 ± 0.019, and non-IID FL was 0.929 ± 0.009, specificity was 0.712 ± 0.204. Non-IID FL attacks achieved recall 0.882 ± 0.088 and specificity 0.712 ± 0.204. SHAP revealed that, in central learning, the most important attack features include MQTT messages, resetting TCP connection, and TCP destination port. Privacy and updating still are the main challenges for deployment, even with Federated Learning (FL).

Keywords—Federated learning; Internet of Things; intrusion detection; explainable artificial intelligence; SHAP; cyber security; privacy-preserving machine learning.

1. INTRODUCTION

Internet connected sensors, actuators, and industrial controls combine to form a massive surface area that connects to the internet. It's even larger and more complicated than most perimeter security systems can handle. This is why new IoT datasets, such as CICIoT2023 and TON_IoT, have been developed to provide researchers with a variety of benign and malicious datasets in order to build and evaluate data-driven Intrusion Detection Systems (IDS) [2], [3]. Edge-IIoTset focuses on centralized and federated evaluation in benchmarking datasets for IoT and IIoT [1]. While such traffic can be further explored to expand edge-based ML IDSs, a centralized approach can potentially create more issues for data governance. In FL, shared models are optimized by participating sites that only communicate model updates with each other [4].

FL makes security analytics in various industries much more compelling than many of the privacy trade-offs that come with centralized approaches. Privacy preserving FL can leverage threat modeling, secure aggregation or differential privacy, and related techniques [13]-[16]. Furthermore, high IDS scores may lead to a false (although, perhaps, secure) sense of confidence (e.g. securely designed protocols). SHAP may be employed to better understand such cases [17]-[19].

There were five components of this study: construction of FL-IDS with representative samples of IoT/IIoT traffic; benchmarking the performances of centralized, IID-FL, and non-IID-FL systems; evaluating the influence of client diversity; explaining federated predictions using SHAP; and discussing the federal-versus-local data tradeoff, but without stating that there is some level of privacy in FedAvg. Four research questions examined whether FL techniques would yield similar results as centralized techniques; how the behavior of non-IID clients would be; what features would determine federated decisions; and what constraints would exist with local raw traffic stays.

2. RELATED WORK

More recent studies on IoT-IDS consider data distribution as a fundamental aspect of the design. In FL-IDS, Campos et al. stated that the evaluation of FL-IDS is particularly dependent on the way IoT data is distributed among the participants and the method of aggregation [5]. On the other hand, Man et al. developed an FL system with an attention-assisted CNN for edge intrusion detection and developed an FL solution [6], while Hamdi focused on centralized and FL-based IoT detection and analyzed the cost of communication [7]. Lastly, Lazzarini et al. analyzed the FedAvg and adaptive federated optimizers for IoT intrusion detection, and illustrated that the choice of the aggregator would depend on the case [8]. This literature supports the need for controlled benchmarking studies in the centralized versus federated for the hypothesis testing of equivalence.

Friha et al. created a safe model exchange system for a decentralized, differentially private IIoT IDS [9]. Yang et al. addressed poisoning-resilient FL for IoT-based intrusion detection [10]. Amiri-Zarandi et al. introduced Social IoT-based trust relationships [11]. Rashid et al. studied FL-based intrusion detection in the IIoT [12]. These studies represent early deployments of FL-IDS. They documented poisoning attacks, non-IID learning, and privacy as major challenges in the deployment of FL-based IDS in the wild [13]-[16].

The same can be said about the research on explainability. In [17], Barnard et al. used SHAP to develop strong network intrusion detection systems. Sharma et al. integrated LIME/SHAP explanations into deep IoT IDS models [18]. In [19], Siganos et al. integrated SHAP into an IoT IDS that is justified. The main challenge now is to develop an integrated analysis of centralized and federated discrimination, the IID/non-IID problem, and explanations at the protocol level with the goal of illustrating the pairwise statistical uncertainty for a real IoT dataset. This work attempts to address this challenge. A very simple MLP is used to ensure that the main differences can be attributed to the training context and the client distribution, rather than to a complex architecture.

3. MATERIALS AND METHODS

Dataset and preprocessing. The authors of Edge-IIoTset provide support for evaluation of federated and centralized learning [1] and also have 14 attacks split into 5 threat groups, 10 IoT devices, and many IIoT/IoT protocols. For these reasons, we selected this dataset. The ML-ready dataset consisted of 157,800 rows and 63 columns, with 133,499 attack labeled rows and 24,301 normal labeled rows with no missing values. We found 814 exact duplicates. We deleted these rows and were left with 156,986 rows. To keep the integrity of our experiment on a CPU, we performed stratified sampling of 60,000 rows. We sampled all 15 attack type classes. To minimize direct identity and memoscopic leakage, we removes IDs, timestamps, IP addresses, high cardinality text fields (including URI/payload strings), and constant fields. We were left with 34 numeric protocol/flow features. For stratification and sampling, we used attack type and created our client to perform the task of predicting whether an example was an attack or normal. Each iteration used an 80/20 stratified split and we only fitted StandardScaler to the training data.

Table 1. Audited dataset and experimental configuration

Characteristic	Verified value
Source file	157,800 × 63
Benign / attack	24,301 / 133,499
Attack types	14 + normal
Missing cells	0
Exact duplicates removed	814
Unique rows after audit	156,986
Analytic sample	60,000
Predictor features	34 numeric
Train/test split	80% / 20%
FL clients / rounds	5 / 12

The first case study used a Modelling and Federated Learning Protocol. Our model consists of binary MLPs with 34 inputs, 2 hidden layers with 32 & 16 ReLU units respectively, and a single output logit. Binary cross entropy loss function was minimized using Adam with Xavier initialization and a learning rate of 0.001. We set batch size to 256. Iteration limit was set to 12. Centralized training was done by pooling together all training records. FedAvg used 5 clients with 12 rounds of communication, each round consisted of a local epoch. Client models were trained in parallel and aggregated after a local epoch. IID clients were assigned attacks in an evenly distributed manner. The non-IID clients were assigned attacks using a Dirichlet distribution ($\alpha = 0.3$) with a minimum of 2,500 training records. In one of the representative seeds (seed 33), the sizes of IID clients were 9,606/9,603/9,601/9,597/9,593, wherein under non-IID allocation, these sizes were 9,422/9,967/10,741/11,092/6,778.

Evaluation and Explainability. Five paired experiments were performed utilizing seeds 11, 22, 33, 44 and 55. Evaluated metrics: accuracy, precision, attack recall, benign specificity, F1, ROC-AUC, false-positive/negative rates, and Matthews correlation coefficient (MCC). For paired F1, AUC and MCC values, two-sided Wilcoxon signed-rank tests were performed. Due to the low number of repetitions (5), we made use of conservative assumptions on p-values and reported the effect direction. As an illustrative case, we interpreted the Federated-IID model in seed 33 with SHAP DeepExplainer and reported the importance values as mean absolute SHAP for attack logit. We made no implementation of differential privacy, encryption or secure aggregation, therefore any privacy claim is limited to data locality. The computational work was done using the following Python libraries: Pub 3.13.5, pandas 2.2.3, scikit-learn 1.8.0, PyTorch 2.10.0, SciPy 1.17.0 and SHAP 0.50.0.

4. RESULTS

Among the five runs, centralized learning provided the strongest balanced discrimination (mean accuracy 0.853 ± 0.013; F1 score 0.906 ± 0.009; ROC-AUC 0.967 ± 0.002; MCC 0.624 ± 0.012; Table 2; Figure 1). In contrast, for the five runs, accuracy was 0.802 ± 0.025, the F1 score was 0.870 ± 0.019, and the AUC was 0.938 ± 0.003. However, the precision was high at 0.981 ± 0.001. Non-IID behaved somewhat differently from being uniformly worse with an accuracy of 0.856 ± 0.044 and an F1 score of 0.910 ± 0.033. This was due to an increase in recall of attacks to 0.882 ± 0.088, where as benign examples were classified as attacking with 0.712 ± 0.204 and a much higher false positive rate of 0.288 ± 0.204. For both attacks, the AUC (0.929 ± 0.009), and MCC (0.556 ± 0.022) were also worse than centralized learning. It was also demonstrated that increased variability meant that in this case MCC and the AUC (0.929 ± 0.009) were less than centralized learning.

Table 2. Performance across five paired repetitions (mean ± SD)

Metric	Central	Fed-IID	Fed-non-IID
Accuracy	0.853±0.013	0.802±0.025	0.856±0.044
Precision	0.987±0.004	0.981±0.001	0.947±0.031
Recall	0.837±0.019	0.781±0.030	0.882±0.088
Specificity	0.941±0.022	0.919±0.008	0.712±0.204
F1	0.906±0.009	0.870±0.019	0.910±0.033
ROC-AUC	0.967±0.002	0.938±0.003	0.929±0.009
MCC	0.624±0.012	0.541±0.030	0.556±0.022

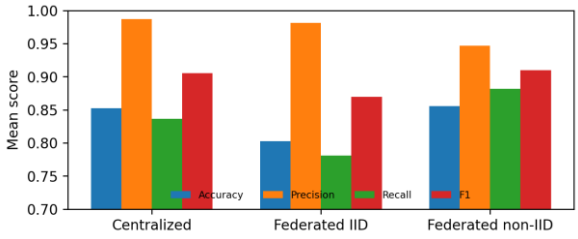

Figure 1. Centralized and federated performance. Source: authors' analysis of Edge-IIoTset.

Table 3 shows that the paired Wilcoxon tests were not significant at the α=0.05 level. For F1 and AUC, p-values were p=0.0625 for both centralized versus IID and centralized versus non-IID. For AUC and MCC, p values were also p=0.0625 for centralized versus non-IID and centralized versus non-IID. From these p values for n=5, we can see consistent directional differences; however, we cannot make a significant inferential claim, and p=0.0625 is not significant. To illustrate and further support this claim, convergence trajectories in Figure 2 show performance for all 12 rounds of communication rather than showing a single final estimate. Finally, local CPU training averages were 8.423s±0.549s for centralized training, 7.537s±0.549s for Fed-IID, and 7.124 for Fed-non-IID. These timings cannot be considered deployment latencies, since they only encompass the local PC and do not consider networking or client availability.

Table 3. Paired Wilcoxon signed-rank comparisons (n=5)

Comparison	W	p	Mean Δ
F1: Central–IID	0	0.0625	+0.0360
F1: Central–non-IID	6	0.8125	−0.0045
AUC: Central–IID	0	0.0625	+0.0291
AUC: Central–non-IID	0	0.0625	+0.0386
MCC: Central–IID	0	0.0625	+0.0836
MCC: Central–non-IID	0	0.0625	+0.0680

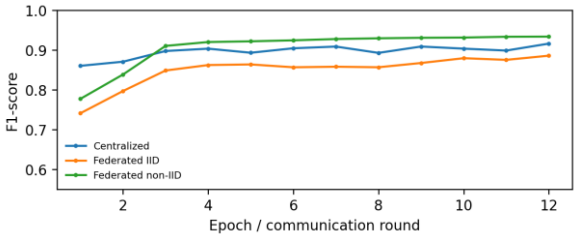

Figure 2. Representative 12-round federated convergence (seed 33). Source: authors' analysis.

SHAP analysis for the Fed-IID model's representative sample showed global features. The TCP destination port feature displayed the greatest (mean of |SHAP| 0.892) influence on classification followed by features MQTT Message Type, TCP Reset, MQTT Header Flag and UDP Stream Identifier (in decreasing order of influence), respectively. TCP ACK flags, aggregate TCP flags, and MQTT Clean Session state influenced the classification. The raw identity classifier was not overly influenced by a single field of the Transport and Application Protocol. This was confirmed by eliminating IP addresses and textual payload identifiers.

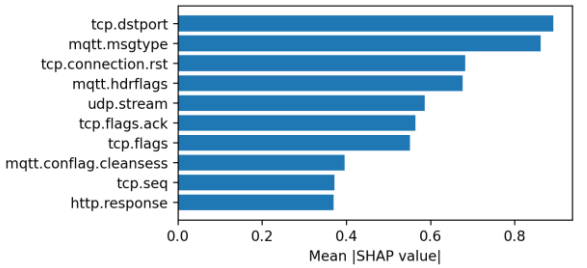

Figure 3. Top global SHAP features for the representative Fed-IID model.

5. DISCUSSION

The system was able to perform meaningful intrusion detection without having to pool training data from the clients. While the system's centralized learning improved balanced performance, it optimized the IID-FL and non-IID-FL the most, where the centralized AUC was improved by approximately 0.029 and 0.039, respectively. Although five-run Wilcoxon tests did not yield p<0.05, these results are considered meaningful. This is consistent with previous studies of FL-IDS which show aggregation and data partitioning significantly affect performance [5], [7], [8]. This also resonates with the need for FL evaluations, to report many metrics other than accuracy.

Non-IID results highlighted RQ2 the most. With heterogenous clients, an overall increase in attack recall and F1 scores was noted, over a large decrease in specificity and more significant general model instability.

Because the focus for Edge-IIoTset is on attack prediction, but at the cost of more benign false alarms, developing models which predict attacks with high precision can also serve to improve accuracy and F1 metrics. In this case, a higher AUC and MCC scores were able to provide more balanced results. For an IoT monitoring systems, high unreliability from false positive predictions can lead to an unnecessary response and decrease the trust of operators. Client-aware calibration and personalized or clustered FL can also be used, along with other forms of aggregation, many communication rounds, and avoiding interpreting a case of non-IID accuracy as stronger generalization.

RQ3 considers explanations. The main SHAP features: TCP Destination Port, MQTT Message Type, TCP Reset, MQTT Header Flags and UDP Stream Behavior, seem to be important protocol-level signals for an IoT/IIoT IDS, and fall within the range of XAI literature that uses SHAP to make intrusion decisions auditable (accessible) [17]-[19]. Explanations do not guarantee causal security mechanisms. SHAP indicates that the model relies on distribution. Therefore, feature importance needs to be assessed across device, time and network shifts. It is essential to evaluate the reasons and mechanisms and thoroughly assess the security mechanisms, especially, prior to implementation.

The wording on privacy in RQ4 was difficult. Centralizing cloud-based raw packet/flow logs in FedAvg is useful for cross-domain collaboration [4], [13], and also reduces collaboration-related exposure. However, training-time data sharing and training-time manipulation of aggregation remain valid security risks [14]-[16]. In this respect, our study focuses on data locality and not privacy. FL should be coupled with secure aggregation, authenticated clients and poisoning attacks, and if tolerable, 2DF-IDS should be coupled, where differential privacy can be applied.

First, one public benchmark was used to build clients rather than using geographically dispersed live networks. Second, the 60,000 records used was repeatedly selected from a large dataset. Third, binary classification was performed rather than classifying attack types. Fourth, only five clients (FedAvg), one MLP, and 12 rounds were implemented. Fifth, statistical inference was performed on five paired repetitions. Communication cost was not claimed, because the local CPU experiment did not construct a real network. SHAP was computed on a representative model and on a subset of the data rather than on each repetition. Auditability was improved, but so was external validity.

Conclusion

This study offers a reproducible approach to analyze explainable and reproducible performance of federated intrusion detection systems using real Edge-IIoT set traffic. After removing duplicates and performing leakage-related feature control, a 34-feature MLP was assessed in a central setting and with 5 clients using FedAvg in an IID and non-IID setting across 5 paired splits. The central learning approach yielded the strongest balanced discrimination (AUC 0.967 ± 0.002; MCC 0.624 ± 0.012). Real-world operational trade-offs were observed in non-IID training. In non-IID training, a higher attack recall was noticed at the expense of a lower and more variable benign specificity. Exact Wilcoxon tests pointed out that no significant differences were observed ($p > 0.05$) after 5 repetitions. Thus, the observed trends in performance are the focus rather than the confirmed significance. SHAP explained the federated model and demonstrated that the auditability of the model was enhanced with no causal relations when the TCP and MQTT protocols were present. Federated learning can lessen the burden of centralizing sensitive IoT data. However, when used on its own, FedAvg is an incomplete privacy mechanism. Further studies are needed on the secure aggregation, differential privacy, robust aggregation, personalized FL, an increase in the client population, multi-class detection, concept drift and real edge-network communication costs.

REFERENCES

[1] M. A. Ferrag, O. Friha, D. Hamouda, L. Maglaras, and H. Janicke, “Edge-IIoTset: A New Comprehensive Realistic Cyber Security Dataset of IoT and IIoT Applications for Centralized and Federated Learning,” IEEE Access, vol. 10, pp. 40281–40306, 2022, doi: 10.1109/ACCESS.2022.3165809.

[2] A. Alsaedi, N. Moustafa, Z. Tari, A. Mahmood, and A. Anwar, “TON_IoT Telemetry Dataset: A New Generation Dataset of IoT and IIoT for Data-Driven Intrusion Detection Systems,” IEEE Access, vol. 8, pp. 165130–165150, 2020, doi: 10.1109/ACCESS.2020.3022862.

[3] E. C. P. Neto, S. Dadkhah, R. Ferreira, A. Zohourian, R. Lu, and A. A. Ghorbani, “CICIoT2023: A Real-Time Dataset and Benchmark for Large-Scale Attacks in IoT Environment,” Sensors, vol. 23, no. 13, Art. no. 5941, 2023, doi: 10.3390/s23135941.

[4] D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, J. Li, and H. V. Poor, “Federated Learning for Internet of Things: A Comprehensive Survey,” IEEE Commun. Surveys Tuts., vol. 23, no. 3, pp. 1622–1658, 2021, doi: 10.1109/COMST.2021.3075439.

[5] E. Mármol Campos, P. Fernández Saura, A. González-Vidal, J. L. Hernández-Ramos, J. Bernal Bernabé, G. Baldini, and A. Skarmeta, “Evaluating Federated Learning for intrusion detection in Internet of Things: Review and challenges,” Comput. Netw., vol. 203, Art. no. 108661, 2022, doi: 10.1016/j.comnet.2021.108661.

[6] D. Man, F. Zeng, W. Yang, M. Yu, J. Lv, and Y. Wang, “Intelligent Intrusion Detection Based on Federated Learning for Edge-Assisted Internet of Things,” Secur. Commun. Netw., Art. no. 9361348, 2021, doi: 10.1155/2021/9361348.

[7] N. Hamdi, “Federated learning-based intrusion detection system for Internet of Things,” Int. J. Inf. Secur., vol. 22, pp. 1937–1948, 2023, doi: 10.1007/s10207-023-00727-6.

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

[9] O. Friha, M. A. Ferrag, M. Benbouzid, T. Berghout, B. Kantarci, and K.-K. R. Choo, “2DF-IDS: Decentralized and differentially private federated learning-based intrusion detection system for industrial IoT,” Comput. Secur., vol. 127, Art. no. 103097, 2023, doi: 10.1016/j.cose.2023.103097.

[10] R. Yang, H. He, Y. Wang, Y. Qu, and W. Zhang,

"Dependable federated learning for IoT intrusion detection against poisoning attacks," Comput. Secur., vol. 132, Art. no. 103381, 2023, doi: 10.1016/j.cose.2023.103381.

[11] M. Amiri-Zarandi, R. A. Dara, and X. Lin, "SIDS: A federated learning approach for intrusion detection in IoT using Social Internet of Things," Comput. Netw., vol. 236, Art. no. 110005, 2023, doi: 10.1016/j.comnet.2023.110005.

[12] M. M. Rashid, S. U. Khan, F. Eusufzai, M. A. Redwan, S. R. Sabuj, and M. Elsharief, "A Federated Learning-Based Approach for Improving Intrusion Detection in Industrial Internet of Things Networks," Network, vol. 3, no. 1, pp. 158–179, 2023, doi: 10.3390/network3010008.

[13] N. Truong, K. Sun, S. Wang, F. Guitton, and Y. Guo, "Privacy preservation in federated learning: An insightful survey from the GDPR perspective," Comput. Secur., vol. 110, Art. no. 102402, 2021, doi: 10.1016/j.cose.2021.102402.

[14] X. Yin, Y. Zhu, and J. Hu, "A Comprehensive Survey of Privacy-Preserving Federated Learning: A Taxonomy, Review, and Future Directions," ACM Comput. Surv., vol. 54, no. 6, 2022, doi: 10.1145/3460427.

[15] K. Zhang, X. Song, C. Zhang, and S. Yu, "Challenges and future directions of secure federated learning: a survey," Front. Comput. Sci., vol. 16, Art. no. 165817, 2022, doi: 10.1007/s11704-021-0598-z.

[16] R. Gosselin, L. Vieu, F. Loukil, and A. Benoit, "Privacy and Security in Federated Learning: A Survey," Appl. Sci., vol. 12, no. 19, Art. no. 9901, 2022, doi: 10.3390/app12199901.

[17] P. Barnard, N. Marchetti, and L. A. DaSilva, "Robust Network Intrusion Detection Through Explainable Artificial Intelligence (XAI)," IEEE Netw. Lett., vol. 4, no. 3, pp. 167–171, 2022, doi: 10.1109/LNET.2022.3186589.

[18] B. Sharma, L. Sharma, C. Lal, and S. Roy, "Explainable artificial intelligence for intrusion detection in IoT networks: A deep learning based approach," Expert Syst. Appl., vol. 238, Art. no. 121751, 2024, doi: 10.1016/j.eswa.2023.121751.

[19] M. Siganos, P. Radoglou-Grammatikis, I. Kotsiuba, E. Markakis, I. Moscholios, S. Goudos, and P. Sarigiannidis, "Explainable AI-based Intrusion Detection in the Internet of Things," in Proc. 18th Int. Conf. Availability, Reliability and Security (ARES '23), 2023, Art. no. 53, pp. 1–10, doi: 10.1145/3600160.3605162.

[20] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in Proc. 20th Int. Conf. Artificial Intelligence and Statistics (AISTATS), PMLR, vol. 54, pp. 1273–1282, 2017.

[21] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Advances in Neural Information Processing Systems 30, pp. 4765–4774, 2017.