Comprehensive Analysis over Centralized and Federated Learning-based Anomaly Detection in Networks with Explainable AI (XAI)

Yasintha Rumesh*, Thulitha Theekshana Senevirathna†,
Pawani Porambage‡*, Madhusanka Liyanage†‡, Mika Ylianttila‡
* VTT Technical Research Centre, Finland
†School of Computer Science, University College Dublin, Ireland
‡ University of Oulu, Finland
Email: *[firstname.lastname]@vtt.fi, †thulitha.senevirathna@ucdconnect.ie, madhusanka@ucd.ie, ‡[firstname.lastname]@oulu.fi

***Abstract*—Many forms of machine learning (ML) and artificial intelligence (AI) techniques are adopted in communication networks to perform all optimizations, security management, and decision-making tasks. Instead of using conventional black-box models, the tendency is to use explainable ML models that provide transparency and accountability. Moreover, Federate Learning (FL) type ML models are becoming more popular than the typical Centralized Learning (CL) models due to the distributed nature of the networks and security privacy concerns. Therefore, it is very timely to research how to find the explainability using Explainable AI (XAI) in different ML models. This paper comprehensively analyzes using XAI in CL and FL-based anomaly detection in networks. We use a deep neural network as the black-box model with two data sets, UNSW-NB15 and NSL-KDD, and SHapley Additive exPlanations (SHAP) as the XAI model. We demonstrate that the FL explanation differs from CL with the client anomaly percentage.**

***Index Terms*—6G, Security, Privacy, Explainable AI, Centralized Learning, Federated Learning.**

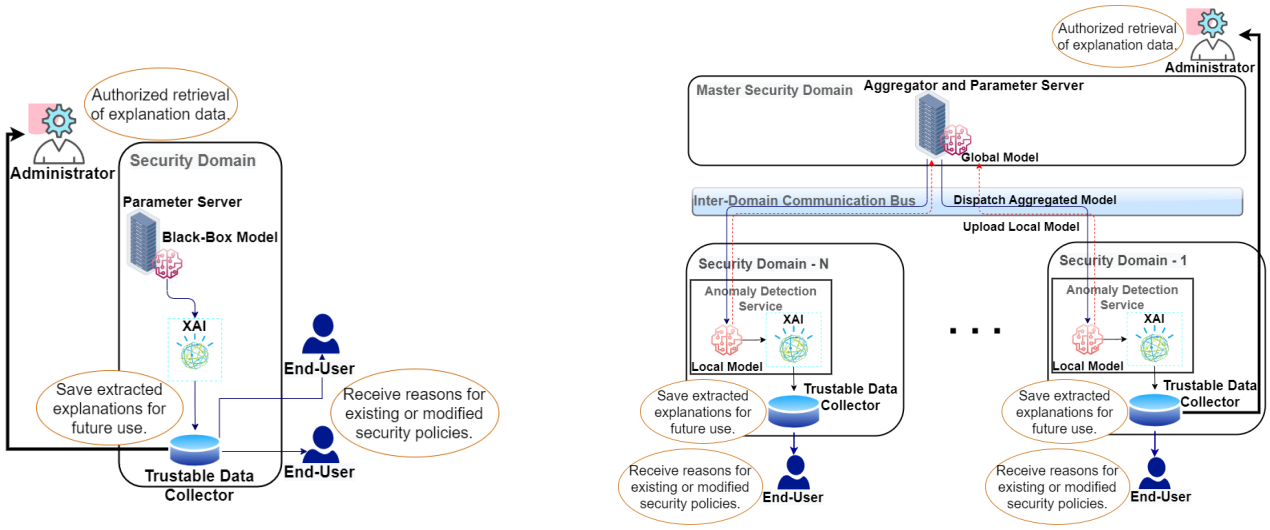
I. Introduction

5G and Beyond 5G (B5G) paves the way to virtualize existing network functions (NFs) which provide differentiated services across several administrative domains achieving guaranteed service performances. Also, it is easier to create adaptable, programmable, and self-managing infrastructures that meet the demanding performance requirements of new and emerging services. Artificial Intelligence (AI) is a key enabler of network automation, lowering operating expenses, increasing productivity, and reducing the risk of human error. Significant security and privacy issues will arise from this proposed ecosystem due to reduced human engagement in service and network administration. With the rise of edge computing capabilities and applications, existing Machine Learning (ML) algorithms face many challenges. More training data is needed in these use cases. Centralized Learning (CL) performance have constrained by cloud processing and storage capacity. Federated Learning (FL) is suggested to be deployed as a secure ML algorithm in future networks. FL outperforms CL in terms of privacy and cost-effectiveness but has yet to attain the needed accuracy in comparison [1]. The challenge is understanding how the FL model behaves and building trust in its decisions.

Network anomaly detection is a crucial area of interest in developing future networks. Researchers are focusing on using FL-based mechanisms to detect anomalies in networks. However, the growing interest in utilizing AI in anomaly detection should pay attention to the possible drawbacks and risks, including threats in FL, data integrity, and AI models. Because of its reasonably accurate and autonomous detection, anomaly detection using ML techniques based on Deep Neural Networks (DNNs) is becoming increasingly popular. These black-box AI systems cannot explain the significant features that affect behavior or the effect of change in input features on the outcome. By incorporating explanation techniques with AI, the trust between the domain expert and AI algorithms can be increased. The automation of future networks requires understanding this discrepancy between AI's potential and its practical applications.

Using unique or modified ML techniques Explainable AI (XAI) adds explainability to black-box models. Explainability will enable the end users to trust the black-box model decision using XAI's practical explanation approaches. This is a primary objective of XAI [2]. It is also possible to use XAI output to analyze the trained model and improve it further. There are comprehensive surveys carried out in integrating XAI into future networks while bridging the gap between them, outlining the roles XAI plays in the network, and demonstrating how to integrate XAI [3]–[6]. It is evident from these surveys that it is vital to integrate explainability into anomaly detection to improve the trust and accountability of AI/ML models. Different stakeholders which interact with multiple verticals and services may require an explanation of how and why the AI/ML models take certain decisions.

We aim to compare how explainability changes in a black-box FL environment. Thus, we used SHapley Additive exPlanations (SHAP) as the XAI model to compare the explainability between CL and FL. We used DNN as the black-box model and the datasets UNSW-NB15 and NSL-KDD. We showed



(a) Explainable Centralized Model

(b) Explainable Federated Model

Fig. 1: Anomaly detection models added with explainability layer.

that the FL explanation differs from the CL explanation, and this difference is higher with the anomaly percentage in the network. We observed that the top features in the SHAP summary plot show a linear relationship with the percentage difference between CL and FL for the feature Standard Deviation (SD) and the percentage difference between CL and FL for the Shapley value. This was accurate for considered attack categories.

The rest of this paper is organized as follows. Section II presents related work about XAI, FL-based anomaly detection, and XAI-based anomaly detection frameworks. Section III describes the explainable models of CL and FL anomaly detection. Section IV describes the experimental setup and results obtained from NSL-KDD and UNSW-NB15 datasets. Sections V and VI compare results and concludes the paper.

II. Related Work

There has been a considerable surge of research in recent years toward improving the explainability of AI models. As depicted, the current state-of-the-art design of XAI models can be categorized into two main aspects;

(1) **Model-based XAI:** build interpretable ML models restricting complexities of such algorithms such as Linear regression, Logistic regression, Decision Trees (DTs), Naive Bayes, k-Nearest Neighbors.

(2) **Post-hoc XAI:** derive explanations for complex ML models by applying methods to analyze the model after training such as Local Interpretable Model-agnostic Explanations (LIME) [7], SHAP [8], Counterfactual Explanations (CE) [9], Layer-wise Relevance Propagation (LRP) [10].

It is important to note that these XAI techniques offer different explainability modes that can be employed appropriately, following the expectations of various stakeholders.

Numerous studies have been conducted on enhancing the efficacy of data-driven anomaly detection algorithms, FL performance enhancement, and techniques for enhancing FL model interpretability. Zhao *et al.* [11] proposed a Multi-Task Deep Neural Network in Federated Learning (MT-DNN-FL) to simultaneously handle the tasks of network anomaly detection, anonymous traffic recognition, and traffic categorization. Since the shared layers can cut back on the number of network parameters, this suggested solution is more effective and superficial. When compared to several single-task models, this multi-task technique can cut down on training time overhead and give domain experts more information regarding network anomalies. Mothukuri *et al.* [12] have presented a FL based on gated recurrent units (GRUs) models for AI-enabled anomaly detection in IoT networks. The accuracy of the global ML model is improved by using the ensemble learning method, which combines updates from several sources. In terms of protecting user data privacy and offering the best accuracy rate for attack detection, the proposed method has excelled over the non-FL variants. Jayasinghe *et al.* [13] have proposed a FL based model incorporating the Zero-touch Network and Service Management (ZSM) architecture for anomaly detection. This is a multi-stage anomaly detector composed of DNNs that performs exceptionally under different compositions of anomaly percentage.

Haffar *et al.* [14] introduced the Random Forests algorithm as a surrogate model to explain FL black-box model misbehavior. Based on updated statistics, this method surpasses existing attack detection systems and has demonstrated high detection rates on several attacks. By identifying the attributes that have been affected by the attack and being able to describe how such attacks work on the peers' end, the surrogate model has increased the explainability of the system. Carletti *et al.* [15] have developed the Depth-based Isolation Forest Feature Importance (DIFFI) algorithm to assess the importance of features for the Isolation Forest (IF) ML model providing a straightforward Root Cause Analysis (RCA). This evaluation method improves the explainability of anomaly detection by allowing stakeholders to gain a profound understanding of

TABLE I: List of Symbols

Symbol	Definition
$\mathcal{F}$	Set of all features
F	A particular feature
S	Subset of features
f_S	Model trained from feature subset S
$f_{S\cup\{F\}}$	Model trained from feature subset $S \cup \{F\}$
x_S	Input with only features from the subset S
$x_{S\cup\{F\}}$	Input with only features from the subset $S\cup\{F\}$
$f_S(x_S)$	f_S model prediction of input x_S
$f_{S\cup\{F\}}(x_{S\cup\{F\}})$	$f_{S\cup\{F\}}$ model prediction of input $x_{S\cup\{F\}}$
$\phi(F)$	Shapley value for feature F
N	Number of features
k	Rank in the SHAP summary plot
$p\%$	Anomaly percentage in local data
$Rank_{CL}^{SHAP}(F)$	The rank of the feature F in the CL SHAP summary plot
$Rank_{Anomaly\_p}^{SHAP}(F)$	The rank of the feature F in the SHAP summary plot for the client with anomaly percentage p
$\sigma_{Anomaly\_p}(F)$	SD of the feature F in FL client data with anomaly percentage p
$\sigma_{CL}(F)$	SD of the feature F in CL data
$\phi_{Anomaly\_p}(F)$	Shapley value of the feature F in FL client data with anomaly percentage p
$\phi_{CL}(F)$	Shapley value of the feature F in CL data

the information given by IF. Huong *et al.* [16] leveraged a Federated learning-based Explainable Anomaly Detection (FedeX) for Industrial Control Systems (ICSs). Variational Auto Encoder (VAE) is used to improve detection efficiency. FL is used to circumvent the lack of centralized training data. SHAP is used to explain the black-box learning model. To the best of our knowledge, this is the only research publication we identified that incorporates XAI in a FL-based anomaly detection technique. This framework improves the model's interpretability while achieving excellent detection performance, learning new data patterns quickly, and being lightweight. On ICSs-weak edge devices, this hybrid model can function effectively. Anomalies or threats may be swiftly detected and contained in each distributed zone thanks to the FL architecture, enabling future ICSs to deal with big data produced by various devices. Integrating XAI increases the system's reliability, enabling specialists to assess and react to abnormalities in distributed ICS environments proactively.

The existing literature shows that there needs to be research on the behavioral change of explainability from CL to FL. It is a vital aspect to focus on with the emergence of FL-based anomaly detection in future networks. It is essential to identify false alarms in anomaly detection systems. Adequate explanations will enable the critical assessment of the decisions of the inference engine. Explanations can be used to overcome the autonomy issues in the network so that system owners can identify how and where to deploy AI. Thus we focus on comparing explanations between the two learning methods.

III. Adding Explainability in Centralized and Federated Learning Models

This section describes centralized and federated learning-based explainable anomaly detection models and their applicability in a networking use case. Figure 1a illustrates the explainable anomaly detection framework with CL. As illustrated in the figure, conventional ML algorithms train the model using the data in a trustable data collector. For instance, this data can be security related to the users and events that occur in the network. Once the model is trained, XAI is applied to analyze the model and its predictions to identify and describe the anomalies in the network flows or the violations of the Security Service Level Agreement (SSLA). Then these explanations are saved in the data services for future use of the network administrators. An end-user who requires an explanation for its existing security policies can obtain the saved explanation for the CL model's predictions. After obtaining the required authorization, the system administrator can retrieve the saved explanation data. The administrator can use these explanations to investigate the model's efficiency or network troubleshooting.

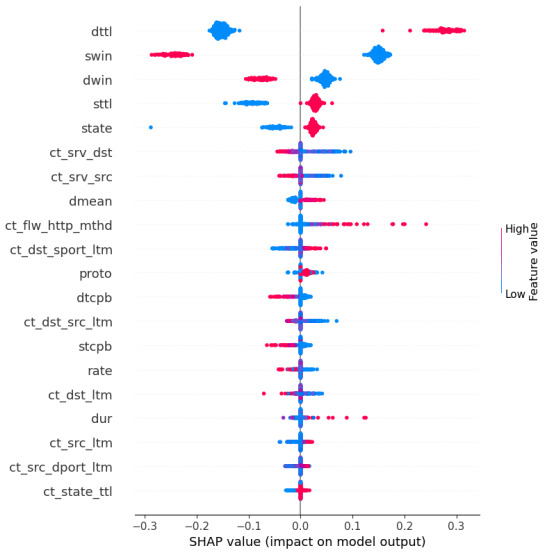


Fig. 2: SHAP summary plot for one client.

The explainable FL-based anomaly detection model is depicted in Figure 1b. Accordingly, the parameter server is placed in the master security domain, whereas other security domains act as FL clients/workers. These are the security services deployed at different locations in the network, which can be considered security domains. The objective is to use FL to detect anomalies in each security domain. This is done by anomaly detection service function. It contains the local model, whereas each security domain interacts with the master security domain via the inter-domain communication bus. The parameter server and anomaly detection service use the inter-domain communication bus to exchange model updates. Anomaly detection service is responsible for deriving insights and predictions based on data collected and entities of the domain. Each of the security domains contains a trustable data collector. These data sources are not visible to the master security domain. It is the reason for integrating XAI in the anomaly detection service with the local model. Integration is done using a post-hoc XAI model since it will be easier to fit in over any black-box model. However, this mechanism can also be applied to model-based XAI, replacing the FL local and

TABLE II: Performance comparison of trained CL and FL models

Dataset	Learning Model	Precision (%)	Recall (%)	F1 score (%)	Accuracy (%)
UNSW-NB15	CL	95.66	94.48	95.07	93.75
	FL	86.43	99.41	92.47	89.67
NSL-KDD	CL	98.99	99.03	99.01	99.04
	FL	96.79	96.03	96.41	96.54

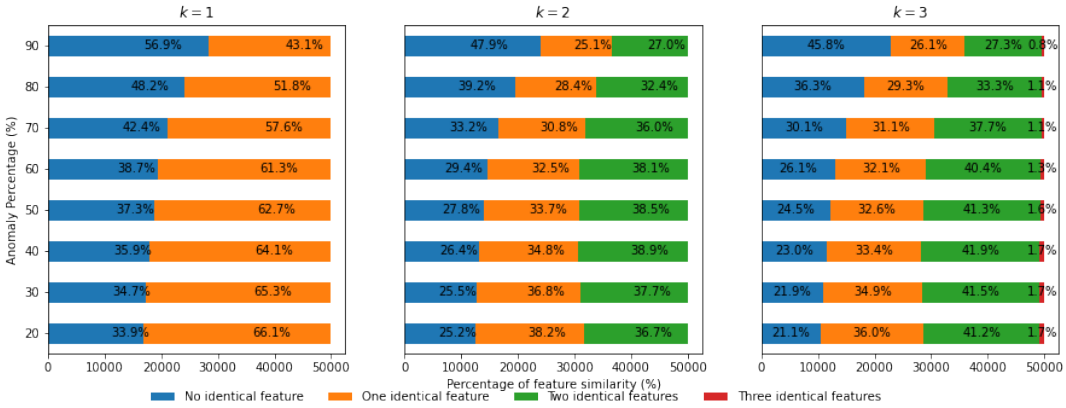


Fig. 3: Comparing identical features between CL and FL explanations in UNSW-NB15 dataset.

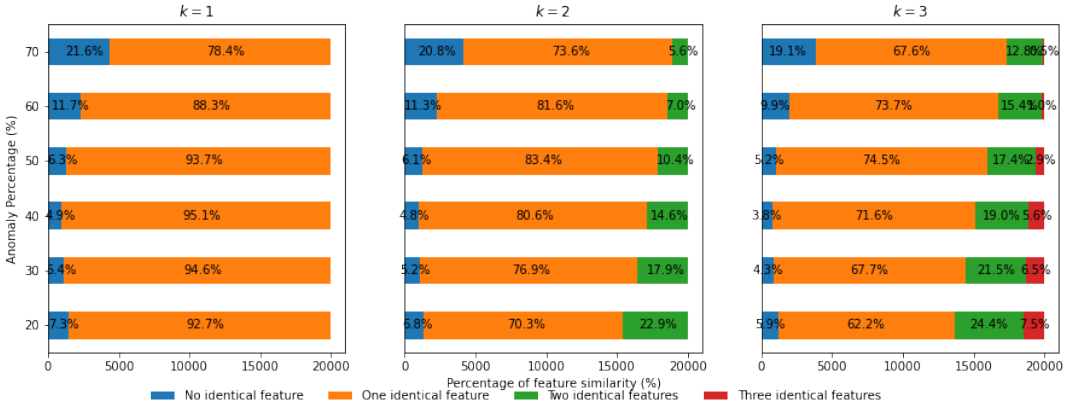


Fig. 4: Comparing identical features between CL and FL explanations in NSL-KDD dataset.

XAI models with interpretable FL local models. End-user or the administrator is able to retrieve the explanation for current service policies at any point of the session as the explanations are saved in the data services.

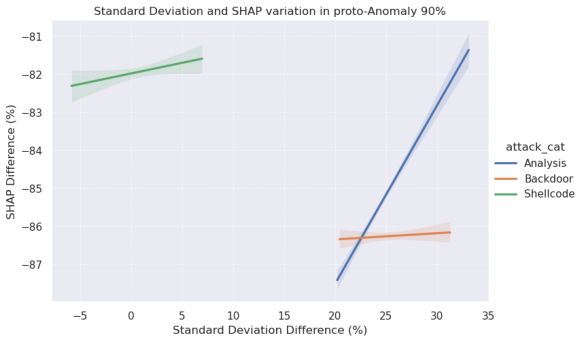
We were interested in generating a global explanation from the train set. Thus, we used SHAP as the XAI tool in the given CL and FL models. SHAP provides a global summary view of the subjected dataset as an explanation. Only some of the XAI algorithms come with this capability. For example, we did not consider LIME for this experiment since it only provided local explanations. Therefore, it is vital to consider the scope of explainability when adopting the XAI model. When auditing for system fairness by the administrator or other external stakeholders, it is beneficial to have an oversight of the system's behavior. The FL model is trained in every iteration using local data in the security domain. XAI would then generate an explanation for each security domain, which is saved in the client's data collector. Furthermore, Table I summarizes the different symbols we used in the paper hereafter.

A. SHAP

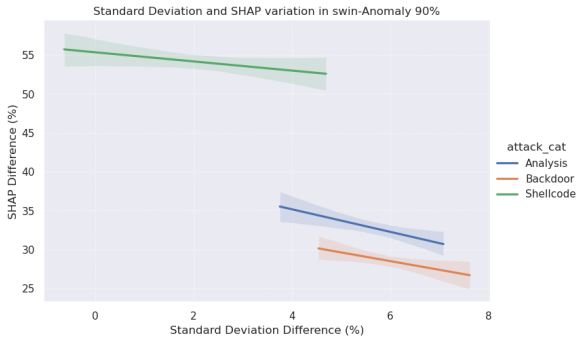
The SHAP method comes from cooperative game theory. It provides a feature importance value called the Shapley value for each input feature. Input features are taken as players. Shapley values give players cooperation to win the game, which is the model outcome. It has a nice characteristic as an expected marginal contribution to the model outcome. If a feature is significant toward the outcome of model prediction, the higher the Shapley value is. $f_S(x_S)$ implies the model prediction of input x_S with only the features of subset S. The missing feature values are replaced with the values from background data. To calculate the Shapley value of feature F, the model outcome is calculated with subset S with feature F as $f_{S\cup\{F\}}(x_{S\cup\{F\}})$. Then the effect of feature F is calculated from the difference between these two model outcomes. Finally calculates the average importance value through perturbation of all feature subsets $S \subseteq \mathcal{F}\backslash\{F\}$.

TABLE III: Performance based on the percentage of anomalies in the client data

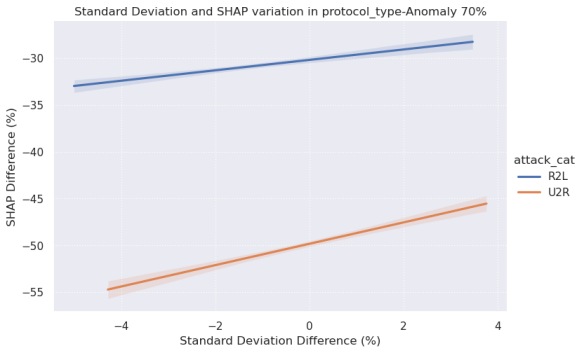
Dataset	Anomaly Percentage	Precision (%)	Recall (%)	F1 score (%)	Accuracy (%)
UNSW-NB15	20%	48.57	99.43	65.26	78.08
	30%	61.91	99.42	76.30	80.87
	40%	71.72	99.42	83.33	83.59
	50%	79.38	99.41	88.27	86.38
	60%	85.45	99.41	91.9	89.17
	70%	90.30	99.41	94.64	91.88
	80%	94.42	99.41	96.85	94.68
	90%	97.75	99.41	98.58	97.35
NSL-KDD	20%	88.31	96.00	92.0	96.82
	30%	93.58	95.97	94.76	96.67
	40%	95.36	95.98	95.67	96.60
	50%	97.03	96.01	96.52	96.50
	60%	98.16	96.04	97.09	96.40
	70%	98.70	96.06	97.36	96.34



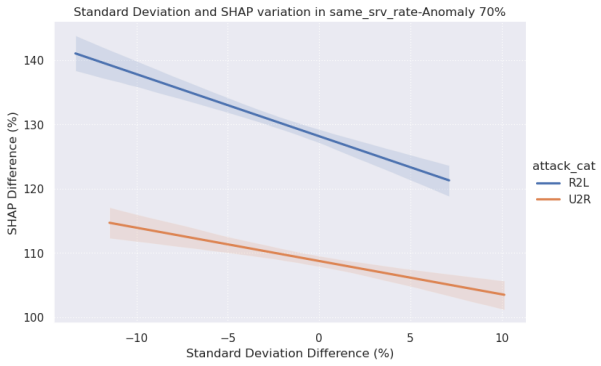
(a) proto feature in UNSW-NB15 for client with anomaly percentage 90%



(b) Swin feature in UNSW-NB15 for client with anomaly percentage 90%



(c) Protocol_type feature in NSL-KDD for client with anomaly percentage 70%



(d) Same_srv_rate feature in NSL-KDD for client with anomaly percentage 70%

Fig. 5: Relationship between SD variation and SHAP value variation as a percentage between CL and FL model.

For a feature F, the Shapley value can be defined as in (1).

$$\phi(F) = \sum_{S \subseteq \mathcal{F} \setminus \{F\}} \frac{|S|!(|\mathcal{F}|-|S|-1)!}{|\mathcal{F}|!} [f_{S \cup \{F\}}(x_{S \cup \{F\}}) - f_S(x_S)] \quad (1)$$

IV. Experiments and Results

This section describes the setup we used to compare explanations between CL and FL models. We used two datasets for model training, testing, and explanation comparison in the experiment. UNSW-NB15 [17], which consists of 257673 flows with 42 features, belongs to 9 different attack categories. NSL-KDD [18] dataset with 148517 flows with 40 features linked with four major attack categories. TensorFlow and TensorFlow-Federated libraries were used to train and test the models. SHAP library was used as the post-hoc XAI model to derive the explanations. The experiment was conducted in three parts: data preprocessing, model training, and deriving explanations. Three categorical features in the used datasets were encoded using the target encoding library in the scikit-learn. Selected 80% of the flows from the datasets randomly to train the models, and the remaining flows were used for testing the models. We used the complete training set for CL model training. This training set was divided between 10 clients for

the FL model training. FL Clients with different anomaly percentages were sampled from the entire training set, ranging from $20\% - 90\%$ in the UNSW-NB15. This was possible since the original dataset contained 64% anomaly percentage. NSL-KDD original dataset had 48% anomaly percentage. Therefore client anomaly percentage varied between $20\% - 70\%$.

Flows in client training samples have been repeated ten times to run several epochs (i.e., small number of local rounds). In the model training phase, the CL model consists of five layers, with three hidden layers. The same keras model was also used in the FL model, which was aggregated using the FedAvg algorithm. FL model was trained with ten rounds. For both cases, adam optimizer was used for compiling the models. In the explanation part of the experiment, Kernel SHAP was used as the explainer as it is more efficient, consistent, and has low bias compared to other existing SHAP explainer types. As the background dataset for the explainer, the entire training set was used in the CL model, and the complete client training set was used for the FL model. SHAP summary plots were generated for each of the attack categories. The CL model provided one global summary plot for one attack category, while the same was generated for each client in the FL model case. This is convenient in FL as the central server does not have access to client data, which is necessary for producing Shapley values. The Shapley value generation for the FL client was done at the client using local data. This experiment was conducted for 100 monte-carlo rounds. Each monte-carlo round included a random sampling of the client dataset with the described distribution. Finally, obtained Shapley values are averaged over all the rounds.

After training ML models for two datasets, the model performance results are shown Table II. Comparatively, models trained from the NSL-KDD dataset showed good classification accuracy. FL model reached an accuracy of 89% and 96% for UNSW-NB15 and NSL-KDD datasets, respectively. In both instances, it is performing less than the CL model.

Figure 2 represents the summary plot achieved for one client at the end of the simulation. This summary plot provides the important features related to the Analysis attack category in that client more elegantly. This summary plot explains how the FL model identified the anomaly type using the behavior of its input features. The top feature affecting the decision of the FL model is the *dttl* feature. Its feature description is the destination to source time to live value. SHAP explains that if the *dttl* value is high, it produces positive Shapley values and obtains the highest absolute Shapley values among other features. Hence it is ranked top in the summary plot. Positive Shapley values indicate that FL model predictions are driving towards classifying that flow as an anomaly and vice versa.

Features ranked order differed from the CL SHAP summary plot to the FL SHAP summary plots. We attempted to summarize results by calculating the percentage of features that are identical up to the k^{th} ranking of the summary plot. At each round, we counted the instances where identical features were observed at the k^{th} ranking of the summary plot for the FL client with anomaly percentage $p\%$ and CL. This can be formulated as $Rank_{CL}^{SHAP}(F) = Rank_{Anomaly\_p}^{SHAP}(F) = k$ and the results for $k = 1, 2, 3$ are shown in the Figure 3 and Figure 4. To explain the ranking order differences in CL and FL summary plots, we evaluated FL model performance with local data. Table III shows the accuracy, precision, recall, and F1 score metrics for local data with different anomaly percentages.

Figure 5 shows how the SD change of a feature from CL to FL affects the Shapley value change. Figure 5 provides the regression plot with confidence intervals for the percentage values in axis calculated as in (2) and (3), respectively.

$$\text{Standard Deviation Difference } (\%) = \frac{\sigma_{Anomaly\_p}(F) - \sigma_{CL}(F)}{\sigma_{CL}(F)} \times 100 \quad (2)$$

$$\text{SHAP Difference } (\%) = \frac{\phi_{Anomaly\_p}(F) - \phi_{CL}(F)}{\phi_{CL}(F)} \times 100 \quad (3)$$

Only SD could show such behavior among other statistical parameters we tested, such as correlation coefficient and mean. Results shown in Figure 5 are for the top two features of the SHAP summary plots for FL clients with the highest anomaly percentage in the network. Figure 5 shows regression lines for various attack categories in the dataset.

V. Discussion

The FL clients provided a different feature ranking order than the CL SHAP summary plot. Figure 3 and Figure 4 indicate that for $k = 1$, there is an excellent chance to have the same feature as the top feature in CL and FL explanations. The percentage of having the same top feature in both CL and FL explanations is significantly lower in the UNSW-NB15 dataset compared to the NSL-KDD dataset. Table III show a significant variation in F1 score and accuracy for UNSW-NB15 clients with different anomaly percentages. This is not the case for the NSL-KDD dataset. The original NSL-KDD dataset contains 48% anomalies.The original UNSW-NB15 dataset contains 64% of anomalies, which causes there to be more anomalous nodes in the network compared to the NSL-KDD setup. This data imbalance has caused UNSW-NB15's performance to deteriorate compared to NSL-KDD. This indicates that even the top features tend to be dissimilar when the network is more anomalous.

Results from Figure 5 showed that an SD change in a feature affects the Shapley value shift differently from feature to feature. This kind of linear relationship was observable only in the top 4 features of the SHAP summary plot and with the clients with the highest anomaly percentage in the network. We believe a high number of anomaly data points helped provide crucial information on the considered attack to identify the linear relationship. The adaptive sampling process in Kernel SHAP with higher dimensional feature sets uses a subset with

lower cardinality to approximate the Shapley values when the feature count is high. This has a significant adverse effect on Shapley value in the lower ranks. Hence, we believe this is why lower-ranked features do not show a similar linear relationship as the top-ranked features.

Typically, SHAP is a computation-intensive method (e.g., for N features, it computes around 2^N subsets). Therefore, computing Shapley values for a more significant number of anomalies in CL is putting strains on the central server's processing capacity and is also time-consuming. Due to the reduced dataset size in FL, this calculation is improved well. Hence we can expect more efficient usage of resources and a real-time explainable FL-based anomaly detection mechanism.

VI. Conclusion

In this paper, we compared SHAP-based explanations for centralized and federated learning models and obtained results using UNSW-NB15 and NSL-KDD datasets. FL clients were created with different percentages of anomalies in the local data. From that, we could show that explanation changed from CL to FL as the anomaly percentage increased. We showed that the most significant features in the summary plot exhibit a relationship between the Shapley value difference and the SD difference of the feature. In future work, we intend to extend our work and improve explainability in FL-based anomaly detection framework for automated security management in ZSM architecture. FL has been the most efficient ML technique to deploy in different network management (i.e., authentication and mobility) functions in Unmanned aerial vehicles (UAVs). Adding explainability for UAV networks is considered future work as well.

Acknowledgment

This work is partly supported by VTT Technical Research Centre of Finland and by Business Finland in SUNSET-6G and DROLO, European Union in SPATIAL (Grant No: 101021808), Academy of Finland in 6Genesis (grant no. 318927) and Science Foundation Ireland under CONNECT phase 2 (Grant no. 13/RC/2077_P2) projects.

References

[1] Y. Liu, X. Yuan, Z. Xiong, J. Kang, X. Wang, and D. Niyato, "Federated learning for 6g communications: Challenges, methods, and future directions," *China Communications*, vol. 17, no. 9, 2020.

[2] D. Gunning and D. Aha, "DARPA's explainable artificial intelligence (XAI) program," *AI Magazine*, vol. 40, no. 2, Jun. 2019. [Online]. Available: https://doi.org/10.1609/aimag.v40i2.2850

[3] S. Wang, M. A. Qureshi, L. Miralles-Pechuán, T. Huynh-The, T. R. Gadekallu, and M. Liyanage, "Explainable ai for b5g/6g: Technical aspects, use cases, and research challenges," 2021. [Online]. Available: https://arxiv.org/abs/2112.04698

[4] W. Guo, "Explainable artificial intelligence for 6g: Improving trust between human and machine," *IEEE Communications Magazine*, vol. 58, no. 6, 2020.

[5] T. Senevirathna, Z. Salazar, V. H. La, S. Marchal, B. Siniarski, M. Liyanage, and S. Wang, "A survey on xai for beyond 5g security: Technical aspects, use cases, challenges and research directions," 2022. [Online]. Available: https://arxiv.org/abs/2204.12822

[6] S. Arisdakessian, O. A. Wahab, A. Mourad, H. Otrok, and M. Guizani, "A survey on iot intrusion detection: Federated learning, game theory, social psychology and explainable ai as future directions," *IEEE Internet of Things Journal*, pp. 1–1, 2022.

[7] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016.

[8] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[9] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the gdpr," *Harv. JL & Tech.*, vol. 31, p. 841, 2017.

[10] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller, "Layer-wise relevance propagation: an overview," *Explainable AI: interpreting, explaining and visualizing deep learning*, 2019.

[11] Y. Zhao, J. Chen, D. Wu, J. Teng, and S. Yu, "Multi-task network anomaly detection using federated learning," in *Proceedings of the tenth international symposium on information and communication technology*, 2019.

[12] V. Mothukuri, P. Khare, R. M. Parizi, S. Pouriyeh, A. Dehghantanha, and G. Srivastava, "Federated-learning-based anomaly detection for iot security attacks," *IEEE Internet of Things Journal*, vol. 9, no. 4, 2021.

[13] S. Jayasinghe, Y. Siriwardhana, P. Porambage, M. Liyanage, and M. Ylianttila, "Federated learning based anomaly detection as an enabler for securing network and service management automation in beyond 5g networks," in *2022 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)*. IEEE, 2022.

[14] R. Haffar, D. Sánchez, and J. Domingo-Ferrer, "Explaining predictions and attacks in federated learning via random forests," *Applied Intelligence*, 2022.

[15] M. Carletti, C. Masiero, A. Beghi, and G. A. Susto, "Explainable machine learning in industry 4.0: Evaluating feature importance in anomaly detection to enable root cause analysis," in *IEEE international conference on systems, man and cybernetics (SMC)*, 2019.

[16] T. T. Huong, T. P. Bac, K. N. Ha, N. V. Hoang, N. X. Hoang, N. T. Hung, and K. P. Tran, "Federated learning-based explainable anomaly detection for industrial control systems," *IEEE Access*, vol. 10, 2022.

[17] N. Moustafa and J. Slay, "Unsw-nb15: a comprehensive data set for network intrusion detection systems (unsw-nb15 network data set)," in *Military Communications and Information Systems Conference*, 2015.

[18] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the kdd cup 99 data set," in *IEEE Symposium on Computational Intelligence for Security and Defense Applications*, 2009.