A sumary of: Federated Explainability for Network Anomaly Characterization

Xabier Sáez-de-Cámara[*][†], Jose Luis Flores[*], Cristóbal Arellano[*], Aitor Urbieta[*], Urko Zurutuza[†]
[*]Ikerlan Technology Research Centre, Basque Research and Technology Alliance (BRTA)
[†]Mondragon Unibertsitatea
Arrasate-Mondragón, Spain
[*]{xsaezdecamara,jlflores,carellano,aurbieta}@ikerlan.es, [†] uzurutuza@mondragon.edu

***Resumen*—Machine learning based systems have shown promising results for intrusion detection due to their ability to learn complex patterns. In particular, unsupervised anomaly detection approaches offer practical advantages as does not require labeling the training data, which is costly and time-consuming. To further address practical concerns, there is a rising interest in adopting federated learning (FL) techniques as a recent ML model training paradigm for distributed settings (e.g., Internet of Things - IoT), thereby addressing challenges such as data privacy, availability and communication cost concerns. However, output generated by unsupervised models provide limited contextual information to security analysts at SOCs, as they usually lack the means to know why a sample was classified as anomalous or cannot distinguish between different types of anomalies, difficulting the extraction of actionable information and correlation with other indicators. Moreover, ML explainability methods have received little attention in FL settings and present additional challenges due to the distributed nature and data locality requirements. We propose a new methodology to characterize and explain the anomalies detected by unsupervised ML-based intrusion detection models in FL settings. We adapt and develop explainability, clustering and cluster validation algorithms to FL settings to mine patterns in the anomalous samples and identify different threats throughout the entire network, demonstrating the results on two network intrusion detection datasets containing real IoT malware, namely Gafgyt and Mirai, and various attack traces. The learned clustering results can be used to classify emerging anomalies, provide additional context that can be leveraged to gain more insight and enable the correlation of the anomalies with alerts triggered by other security solutions.**

***Index Terms*—Federated Learning, Anomaly Detection, IoT Malware, Intrusion Detection, Explainable AI**

Tipo de contribución: *Investigación ya publicada*

I. Introduction

Recent advancements in Intrusion Detection Systems (IDS) leverage Machine Learning (ML) and Deep Learning (DL) for enhanced detection over traditional systems, exploring supervised, unsupervised, and semi-supervised approaches. Federated Learning (FL) emerges as a promising paradigm, offering data privacy and efficiency by training models across distributed clients without sharing their data. However, challenges such as data labeling, interpretability, and integration of Explainable AI (XAI) in FL environments persist, limiting broader adoption.

This work, published in [1] introduces a novel methodology to explain and characterize anomalies in ML/DL-based IDS within a FL setting, utilizing XAI and clustering techniques. Evaluated on real IoT malware data, our approach aims to improve security awareness across federated networks, addressing critical issues in IDS interpretability and effectiveness. The proposed solution's compatibility with security information and event management systems enhances its practical utility in real-world scenarios.

Contributions:

- A new approach to characterize and explain anomalies in FL-based IDS using SHAP (Shapley Additive Explanations) for XAI and federated k-means for clustering.
- Experimental validation on diverse IoT security datasets, showcasing the methodology's efficacy in detecting a wide range of attacks.
- Demonstration of interoperability with security solutions, facilitated by Intrusion Detection Message Exchange Format (IDMEF) for alert message exchange, enhancing anomaly response capabilities.

Source code and details are available at https://gitlab.danz.eus/groups/datasharing/federated-anomaly-detection-iot.

II. Proposed system model

The diagram of all the components involved in the proposed method is shown in Figura 1. The diagram is divided into three main blocks: (i) anomaly detection model training, (ii) model inference and (iii) explainer model training and the characterization of the anomalies.

The main focus of this manuscript is not on the FL anomaly detection model training or inference, but on the third block regarding the FL explainer training and anomaly characterization, as denoted by the steps with a shaded background in Figura 1.

As shown in Figura 1, the last block includes two steps that are performed in a federated way: the explainer model training and the characterization of the anomalies. We will use Kernel SHAP to train the explainer model, which requires two inputs, the prediction model f and a background dataset. The output of this step is the explainer model g. The prediction model f is the global anomaly detection model trained with FL, which is common to all clients. To ensure that all clients have the same explainer model g, the same background dataset must be used, which is usually a representative subsample of the training data. However, the data in FL settings are distributed across all clients and not shared. To generate a common representative background set as a subsample of the entire distributed dataset, we will leverage and adapt a federated version of k-means based on k-FED [2]. In this step, the k from k-means refers to the number of subsampled data samples to be used as the background for SHAP.

The anomaly characterization process is the second step that requires the use of FL. Explanations generated for the

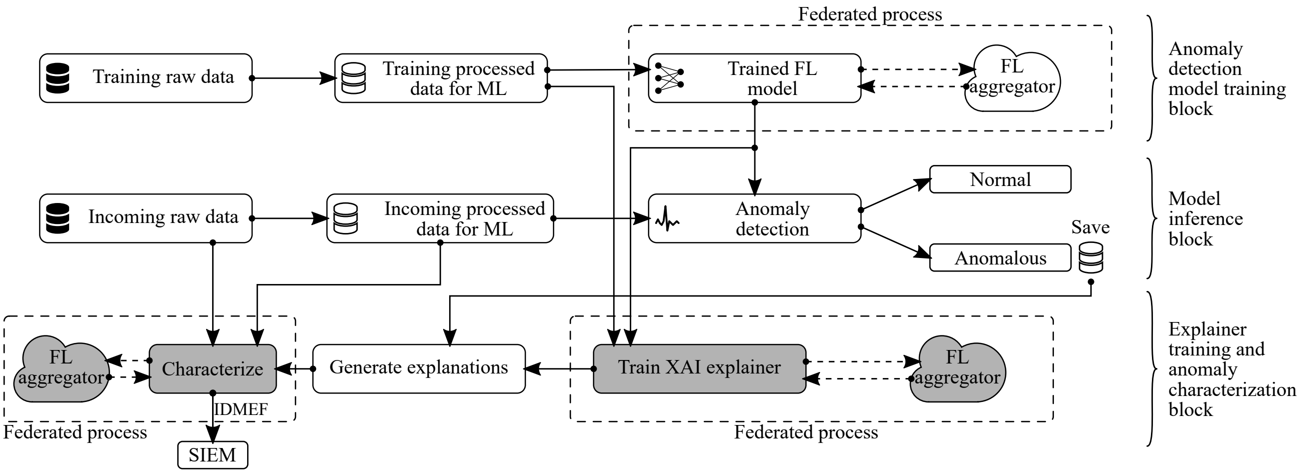

Figura 1. Diagram of the proposed methodology. The components within the dashed frames represent steps performed using FL. Those with shaded background refer to the contributions of this paper.

anomalous samples are inputs of this process, that is, the ϕ_i SHAP values showing the importance of each feature. The other inputs are the processed data and the raw data of the anomalous samples. Since anomaly explanations are local to each client, we use FL to ensure that all clients are able to know and identify all the different anomalous activities found across the federated network, even if each client has been exposed to a different set of attacks. Specifically, we will leverage k-FED [2] to group the explainability results in each client and share it with other peers in the network so that all can have the same clustering labels to refer to the same anomalous instances. In this step, k refers to the global number of anomalous behaviors found.

III. Results

The methodology explains and characterizes anomalies of unsupervised intrusion detection models in a federated learning setting, where the clients throughout the network can have differences in data or behavior distribution and might also be exposed to distinct types of attacks. The explanations are based on the Kernel SHAP model-agnostic method, using a federated version of the k-means algorithm to subsample the background dataset required for SHAP model training across all the clients. We leverage the generated explanations by clustering (in the SHAP space) all the identified anomalies in the network using again an adapted version of the federated k-means algorithm. Since the number of anomalous patterns or groups is not known a priori, we also propose an adaptation of the Calinski-Harabasz internal cluster validation metric for distributed settings to allow the estimation of a suitable number of anomalous clusters found among all the clients.

A practical benefit of the proposed method is that all the federated steps can be performed in a one-shot manner (a single round of communication), which reduces the data transmission between the clients and the FL aggregation server. However, we note that selecting an adequate number of anomalous clusters requires repeating the federated k-means process for different values of k. Additionally, for robustness, it is recommended to perform various trials for the same k to account for random processes, such as the initialization of the centroids in the k-FED k-means process, as the experimental results show high variability in the Calinski-Harabasz scores. While each process requires minimal data transmission overhead proportional to k, multiple trials and repetitions can rapidly increase the cost; for communication efficiency, this should be considered compared to the amount of local training data on each device.

The proposed method identified several anomalous behaviors in the evaluated datasets and assigned a label to each of them that can be used to identify and characterize groups of anomalies. The labels are shared and known to all the clients and serve as a naming system to refer to the same anomalous patterns across all the clients in the federated network. New incoming alerts can be grouped and auto-labeled into the known anomaly behaviors, which can be used to send contextualized alerts representing multiple anomalies using the IDMEF message format, as shown in the results, for interoperability with third-party tools.

Acknowledgements

The European commission financially supported this work through Horizon Europe program under the IDUNN project (N. 101021911). It was also partially supported by the Department of Economic Development, Sustainability and Environment of the Basque Government under the ELKARTEK 2023 program, project BEACON (KK-2023/00085). Urko Zurutuza is part of the Intelligent Systems for Industrial Systems research group of Mondragon Unibertsitatea (IT1676-22), supported by the Department of Education, Universities and Research of the Basque Government.

Referencias

[1] X. Sáez-de Cámara, J. L. Flores, C. Arellano, A. Urbieta, and U. Zurutuza, "Federated explainability for network anomaly characterization," in *Proceedings of the 26th International Symposium on Research in Attacks, Intrusions and Defenses*, ser. RAID '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 346–365. [Online]. Available: https://doi.org/10.1145/3607199.3607234

[2] D. K. Dennis, T. Li, and V. Smith, "Heterogeneity for the win: One-shot federated clustering," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 2611–2620. [Online]. Available: https://proceedings.mlr.press/v139/dennis21a.html