Towards Transparent AI-Powered Cybersecurity in Financial Systems: The Deployment of Federated Learning and Explainable AI in the CaixaBank Pilot

Aikaterini Karampasi[1], Panagiotis Radoglou-Grammatikis[1], Marek Pawlicki[2,3], Ryszard Choraś[3], Ramon Martin de Pozuelo[4], Panagiotis Sarigiannidis[1], Damian Puchalski[2], Aleksandra Pawlicka[2,5], Rafał Kozik[2,3], and Michał Choraś[2,3]

[1]**Department of Electrical and Computer Engineering, University of Western Macedonia, Kozani, Greece**
[2]**ITTI sp. z o.o., Poznań, Poland**
[3]**PBS Bydgoszcz, University of Science and Technology, Bydgoszcz, Poland**
[4]**Technology and Information Security, CaixaBank, Barcelona, Spain**
[5]**University of Warsaw, Warsaw, Poland**

ABSTRACT

In the domain of financial cybersecurity, where trust and reliability is paramount, the advent of Artificial Intelligence is bringing novel tools for network intrusion detection. This paper introduces AI4FIDS, a novel AI-powered Intrusion Detection System leveraging Federated Learning (FL) to enhance data privacy while enabling decentralized model training across multiple financial entities. Concurrently, we present TRUST4AI.XAI, an explainability module designed to render AI decision-making transparent and interpretable, thereby aligning with the critical need for model accountability in financial applications. Our experimental results, conducted in the framework of the AI4CYBER project's financial sector pilot, demonstrate in detecting network intrusions in financial infrastructure while maintaining user privacy, while increasing trustworthiness via explainability methods. The integration of these technologies addresses the dual challenges of effective threat detection and regulatory compliance, offering a scalable solution for modern financial institutions. This work contributes to the ongoing dialogue on leveraging AI for financial security and sets a benchmark for the development of privacy-preserving, interpretable AI models in this sector.

Keywords: Federated Learning; Network Intrusion Detection; Fintech; AI explainability

INTRODUCTION

Cybersecurity plays a crucial role in protecting sensitive information and assuring the continuity of critical sectors' operations in today's interconnected world. It is particularly important in sectors such as the financial domain and banking services, in which trustworthiness and reliability are essential to preserve social trust and economic stability, and which are targeted by a variety of malicious actors [25]. As evolving cyber threats target critical sectors with more and more sophisticated attack attempts, investing in advanced security measures, often assisted by Artificial Intelligence (AI) capabilities, becomes essential to maintain the robustness and reliability of key services. Digitalization of the financial sector, the growth of fintech as a whole, as a part of international critical infrastructure [12] and its vulnerability due to the sensitive nature of financial data are reasons why the financial sector became one of the prime targets of cyber criminals [32]. These are also reasons for increased frequency and impact of malicious attempts targeting financial systems over the last years [2].

At the same time, recent advances in AI and successful applications of machine learning (ML) in detecting and classifying intrusions at the network level have made AI widely recognized as a major tool for enhancing the cybersecurity of banks. However, the use of AI-assisted cybersecurity poses new challenges, such as issues of model transparency and inherent ML/AI vulnerabilities [2].

In this paper, a trustworthy network intrusion detection pipeline is proposed. It includes AI4FIDS – an AI-powered Intrusion Detection System (IDS) which leverages Federated Learning (FL), enabling the training of federated models across multiple decentralized entities or environments. The TRUST4AI.XAI tool providing explainability mechanisms for AI models is also introduced. It is used in conjunction with AI4FIDS to make cybersecurity decisions proposed by the AI-powered system more transparent and interpretable. Both tools are part of a wider architecture, designed, developed and implemented in the AI4CYBER [AI4CYBER] project, which is a European Union (EU) project co-funded by the Horizon Europe research and innovation programme under the Grant Agreement (GA) No 101021936. AI4FIDS is one of the core components of the AI4CYBER project together with tools for root cause analysis, attack simulation, fixing and testing, vulnerability analysis and many more. The proposed explainability module is included in the framework providing trustworthiness for AI services developed in the project. AI4CYBER tools are validated in three real-world pilots. For the purpose of this work, the focus is on the financial sector.

This paper outlines several significant advancements in AI-powered cybersecurity deployments for financial systems through a detailed exploration of the CaixaBank pilot. Major contributions include:

- A novel AI-powered IDS that utilizes Federated Learning to ensure data privacy across multiple financial entities while enabling decentralized model training
- An explainability module that serves an array of xAI methods, giving the user a look at the AI model from different perspectives, contributing to the transparency and interpretability of AI decision-making, thus meeting the need for model accountability in financial applications
- Provides the Experimental Validation of the end-user-centric and sector-oriented AI4FIDS and TRUST4AI.xAI deployments in the CaixaBank pilot.
- Sets a precedent for the development of privacy-preserving, interpretable AI models in the financial sector

The paper is organized as follows: Section II provides the current state of the art related to federated learning in cybersecurity and to explainable AI (xAI). Section III details the approach and design of AI4FIDS and TRUST4AI.XAI solutions and CaixaBank use-case in which we validate the proposed tools. Section IV focuses on the experiments and experimental results obtained to prove the effectiveness of the proposed suite of tools in the banking pilot. Section V concludes this article.

STATE-OF-THE-ART AND RELATED WORKS

Federated Learning for Network Intrusion Detection

The impact of Federated Learning in cybersecurity has been considered by a variety of research studies, with particular focus on its application in intrusion detection and prevention. In the following paragraphs, an overview of relevant survey papers in this field will be discussed. More notably, M. Alazab et al. [3] evaluated the manner in which FL operates and contributes in the context of cybersecurity, with a targeted analysis of selected use case scenarios, applications and confrontations. In the same manner, B. Ghimire and B. Rawat [15] explore the progression of FL and cybersecurity in a reciprocal fashion. The authors initially investigate the utilization of FL in cybersecurity applications, including but not limited to intrusion detection, with peculiar interest on Internet of Things (IoT) and Cyber-Physical Systems (CPS), while they subsequently discuss the impact of cybersecurity in FL. In an extensive review paper, E. M. Campos et al.[8] investigate the manner in which FL is employed for IoT environments, exploring the effect of FL in intrusion detection while also considering the progression of Machine Learning and Deep Learning (DL) approaches by reason of FL. Eventually, areas with potentials for additional exploration and avenues for future studies are described. A comprehensive survey is provided by L. Lavaur et al. [19] regarding the evolution of federated IDS and Intrusion Prevention System (IPS). After establishing their methodological approach, the authors conduct a detailed analysis of the existing research, evaluating a range of criteria including the following: detection techniques, mitigation tactics, data sources and datasets, variations of FL, local models and aggregation methodologies, as well as communication protocols (e.g. overhead optimisation and encryption procedures). In light of the aforementioned criteria, a pertinent categorization of federated IDS and IPS is proposed, and a comparative analysis of the subject literature is undertaken. Subsequently, the authors present a discussion of the open issues and research directions that remain to be addressed. S. Arisdakessian et al. [7] conducted a survey regarding intrusion detection with respect to IoT applications. The authors combined and discussed a variety of technological and research areas, including FL, game theory, social psychology and Explainable Artificial Intelligence. A total of 19 criteria were taken into consideration in order to conduct an exhaustive study and analysis of several works which allowed the identification of significant research gaps pertaining to the aforementioned technological and research areas.

A federated DL framework is presented by S. I. Popoola et al. [26] which consists of multiple distinct nodes performing the training procedure of Deep Neural Networks (DNNs) based on the data provided by their local network traffic. Subsequently, a central server gathers the disparate parameters of each trained model and aggregates them using the Fed+ fusion technique, eventually distributing them back to the nodes. With respect to the DNNs' designation, these incorporate the input layer, two fully connected hidden layers and the output layer. The results that were attained, based on simulations that were executed, provided the authors with an accuracy of 99.27%, precision 97.03%, TPR 98.06%, and an F1-score of 97.50%. This outcome indicates better performance of the federated DL over the local DNN models. With regard to the identification of the optimal fusion technique, a variety of methods were employed, namely Federated Averaging (FedAvg), Fed+, and Coordinate Median (CM). The experiments indicated that Fed+ exceeded the performance of the other state-of-the-art (SOTA) methods, providing evidence for the overall superiority of the DNN-Fed+ model using FL for the intrusion detection assignment in heterogeneous wireless networks.

T. Dong et al. [11] proposed a novel intrusion detection system based on a learning-based methodology, namely FedForest, encompassing FL and Gradient Boosting Decision Trees (GBDT). The proposed technique is implemented by training a local encoder (GBDT classifier) on the distinct clients. The data based on which the clients were trained, were distinct private datasets of each client, while the parameters that were attained in each case were broadcast to the server. Consequently, the server decides for the finest encoders and transmits them to all clients. Eventually, the clients utilize the encoders to encode their data, train and deploy the new models. To further enhance data privacy, a random masking algorithm was utilized on the data. During the evaluation procedure, an illustration of the superiority of the proposed FedForest was performed with a Multi-layer Perceptron (MLP) composed of 3, 5, and 7 layers. The results that were attained indicated accuracy levels of 67.03% on the DDoS2019 dataset, 89.63% on MalDroid2020, 86.76% on Darknet2020, and 79.6% on DoHBrw2020, signifying the prevalence of the suggested methodology.

P. H. Mirzaee et al. [22] suggested a Federated Intrusion Detection System (FIDS) methodological scheme specifically implemented for 5G environments, the primary goal of which was to establish user privacy while simultaneously preserving a high detection rate. More notably, a federated DNN implementation was proposed, appropriate for ensuring privacy of the user's information. The algorithm encompassed a dedicated server for the aggregation of the updates from each respective local model, while the obtained results were sent back to the end nodes. Regarding the evaluation procedure, it was exhibited that the recommended implementation accomplished 99.5% in all metrics, namely accuracy, precision and F1-score, on the NSL-KDD dataset.

W. Schneble and G. Thamilarasu [29] proposed a widely distributed IDS based on ML methodologies, and more specifically they employed FL techniques for Medical Cyber-Physical Systems (MCPS), towards reducing communication and computation costs while increasing network security. The suggested model was evaluated on both real and simulated attacks such as Denial of Service (DoS), Data Modification, and Data Injection. The results that were attained showcased that the model under discussion outperformed SOTA methodologies by achieving 99% accuracy levels and an FPR of 1%, while simultaneously the communication costs were decreased.

O. Aouedi et al. [5] suggested the FLUIDS which describes a semi-supervised implementation for IDS, composed of encoders and trained on each end device with unlabeled data. The local models are afterwards aggregated to be trained on labeled data, which is located on a server, in a supervised manner, eventually providing an ameliorated classification of attacks. B. Li et al. [20] introduced a federated IDS especially trained on detecting DDoS attacks based on prototypical features extracted by GRU layers to eventually derive 97% accuracy. On the other hand, R. Zhao et al. [33] implemented an FL architecture based on BiLSTM towards identifying high-risk malicious behaviour, which had minor variations from a centralized model, to obtain 99.21% accuracy. The IoTDefender was proposed by Y. Fan et al. [13] for 5G IoT through a federated transfer learning architecture which surpassed the performance of traditional implementations achieving 91.93% accuracy and finer generalization abilities. O. Friha et al. [14] proposed the FELIDS framework, which was based on CNN and DNN architectures towards constructing an FL-based IDS model, outperforming other centralized architectures in maintaining privacy of the utilized data as well as high detection accuracy.

Explainable AI Methods for Network Intrusion Detection

Many of the best-performing AI/ML methods function as black boxes, which presents significant ethical concerns for their use in various domains. This lack of transparency can undermine trust and become an obstacle for numerous potentially beneficial applications [9]. This creates a need for reliable interpretability methods, which would facilitate a way for the human operator to understand the decision-making process of the model. With this pressing necessity, xAI is now an intense area of research, with numerous emerging approaches [24].

In this work, the focus is on providing xAI-derived explanations, relying on the representation of AI models using methods that are easier to interpret. These methods can be derived from the original AI models or built from scratch using the available data. The techniques can be broadly classified into two categories: model-agnostic and model-specific methods [27]. The methods that are model-agnostic can be used for any ML model. The tool deployed in the project provides a plethora of xAI methods, including LIME, SHAP, DiCE and ProtoDash.

LIME (Local Interpretable Model-Agnostic Explanations) generates locally faithful explanations by fitting an interpretable model to the neighborhood of the input data, providing insights into complex models by creating simpler, locally interpretable linear models [27].

SHAP (SHapley Additive exPlanations) assigns a value to each input feature based on its contribution to the model's prediction using Shapley values to fairly allocate the payoff among features [21].

DiCE (Diverse Counterfactual Explanations) offers counterfactual explanations by synthesizing data points by perturbing features of a sample until the label flips, essentially presenting a 'what-if' scenario [23].

ProtoDash identifies 'Prototypical Samples' within a dataset to gain insights into the characteristics of a subset or specific class of data through its most representative samples. The samples are found by maximising similarity for a concise representation [18].

The ANCHORS explainer identifies "anchors," or rule-based conditions that reliably predict the same outcome when met. It reveals key factors influencing a model's decisions by testing feature combinations to find those that consistently lead to the same prediction. ANCHORS is model-agnostic [28].

PDP (Partial Dependence Plot) is a visualization tool that shows how a feature impacts the predicted outcome, on average, across a dataset. It highlights the global effect of a single feature, ignoring interactions with others, and helps interpret complex models by offering a clear graphical view of feature influence on predictions [17].

ICE (Individual Conditional Expectation) plots show how predictions change for each individual instance as a feature is varied, revealing interactions and variability in the model's behaviour for specific data points [16].

ALE (Accumulated Local Effects) plots calculate a model's predictions within intervals of a feature, accumulating these effects across the feature range. This offers a realistic view of how changes in the feature influence predictions [6].

Permutation Feature Importance (PFI) is a simple, model-agnostic method that measures the impact of each feature on model performance by shuffling its values. By disrupting each feature and observing the change in accuracy, PFI provides clear insights into feature relevance, making it broadly applicable and easy to interpret, regardless of the model structure [4].

AI-EMPOWERED ANOMALY DETECTION IN BANKING SCENARIO

AI4FIDS - Anomaly Detection Tool

AI4FIDS is illustrated through the C4 model which demonstrates the architecture of the proposed implementation. In general, the C4 model is a structural representation extensively employed in software engineering for conceptualizing and substantiating the architecture of the software systems. Context, Containers, Components, and Code are the pillars of this model, proposed by Simon Brown, to substitute the distinguishable levels of abstraction in the model, each of which offers a distinct perspective of the system, thus making it more feasible to comprehend and explain the architecture to the involved stakeholders, both technical and non-technical. More specifically, Fig. 1 depicts the Context level of AI4FIDS, illustrating its communication with the other entities, interior and exterior.

To begin with, as an IDS based on multiple data sources, AI4FIDS gathers its input from captured network traffic, system logs and operational data which are captured in the Critical System under inspection by AI4FIDS. This data originates from the connections of the Critical System with the End Users and/or External Networks/Systems (i.e., the Internet). The purpose of AI4FIDS is to analyse this data and identify pertinent cyberattacks and anomalous behaviour. Upon producing its results, AI4FIDS spreads the equivalent security events to the Security Information and Event Management (SIEM), which is an external system. SIEM is mainly responsible for improving the security posture of an entity by provisioning for real-time perceptibility regarding security incidents and threats, hence assisting in the regulations' and policies' compliance actions of the entity. More notably, the core concept behind a SIEM system is to normalize, prioritize and correlate the events provided as input from AI4FIDS, while eventually the System Security Operator is able to supervise, evaluate and determine the AI4FIDS security events analysed by SIEM.

Even though the Context level offers a comprehensive examination of the system's engagement with exogenous entities, the Container level provides a more detailed analysis of the architectural underpinnings of AI4FIDS. More specifically, in the Container level, the communication of the distinct entities, identified as logical, that comprise AI4FIDS are described, while the interaction with extrinsic components is depicted. To that end, the following containers might be identified while constructing the FIDS in AI4FIDS: (a) Log-based (L-FIDS), (b) Operational Data-based (O-FIDS), (c) Network Flows-based (N-FIDS), (d) Visual-based (V-FIDS) and (f) Training for FIDS (T4FIDS). Then, in the architectural representation of the system, on top of the Context and Container levels, the Component level is able to specifically define the architecture of the distinct components along with their communications.

TRUST4AI.XAI - AI Explainer

For the AI explainability purposes, the modular architecture leveraging microservices is proposed for the TRUST4AI.XAI Explainer. The modularity and microservice approach ensures the scalability, maintainability, and flexibility of the solution. The system allows the end user to perform analyses of any supervised learning model with minimal setup. This process is based on communication between the xAI components of the TRUST4AI.XAI system and the AI models, integrated via REST API or via Apache Kafka, depending on use case needs. The entire TRUST4AI.XAI system is an arrangement of microservices, as illustrated in Fig. 2. The xAI serving component features a user-friendly, React-based front-end. The main function of the component, as it is provided by the frontend, is to easily allow the user to perform analyses and visualize the decision-making process undertaken by the AI classification tools in graphical/chart form. It is important to note that the AI models are external to the xAI system. Another component, as seen in Fig. 2, is API Gateway, written using the Spring framework in Java. It serves as a gateway between the frontend and microservices. The API Gateway orchestrates the local and global microservices, the preprocessing microservices, and the data sink. The user, using the web application/frontend can issue requests to explain particular samples for a particular model, choosing from the samples visible on an APACHE Kafka topic. Those samples are the concatenation of the feature vector and the classification results coming from the AI4FIDS component, which are pushed to the xAI Kafka topic.

The microservices are responsible for creating respectively both local and global explanation objects and visualizing their analysis. The interdependencies of these components have been illustrated in Fig. 2. The architecture also includes components such as a service that collects logs from individual microservices, and a configuration centre.

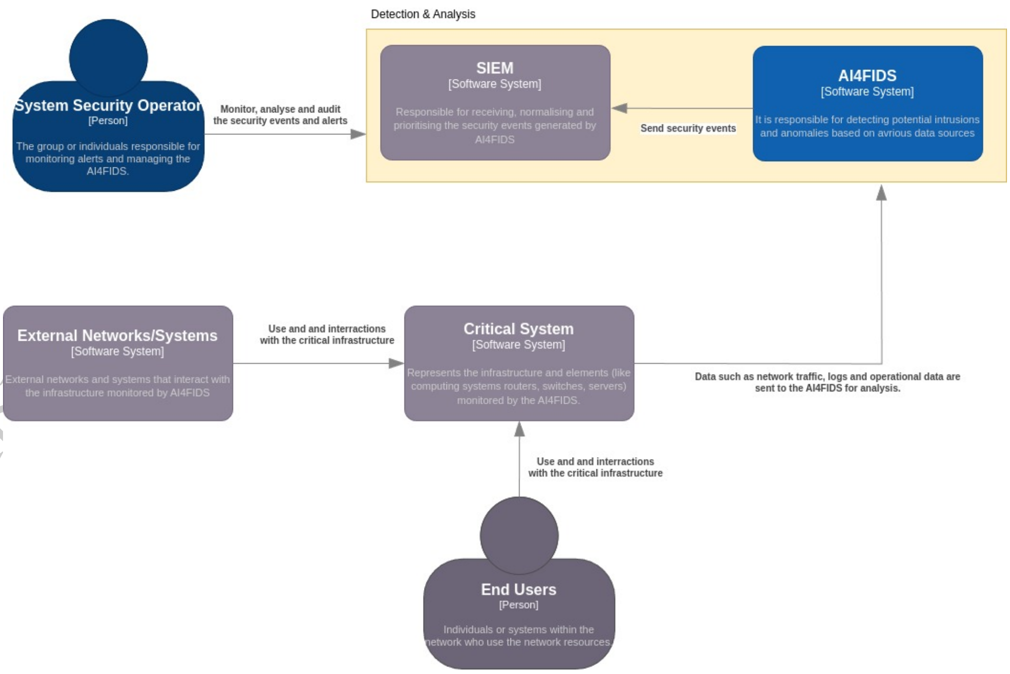

Figure 1. AI4FIDS system context model.

CaixaBank Scenario

The scenario defined for the validation of the proposed pipeline of AI4FIDS and TRUST4AI.XAI solutions is rooted in the AI4CYBER project finance sector cybersecurity use case. The piloting partner is CaixaBank, one of the leading financial institutions in Spain, chosen as the ”Best Bank in Western Europe 2024” [Daly]. CaixaBank employs a set of both proprietary and commercial cybersec solutions. Tools for vulnerability

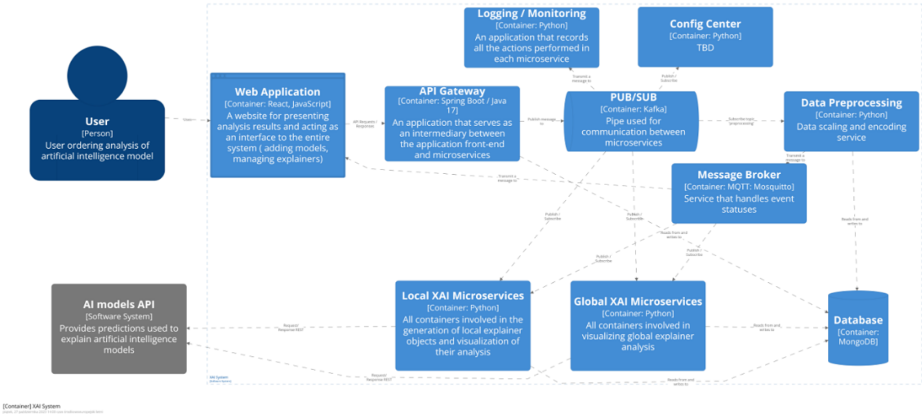

Figure 2. The architecture of the TRUST4AI.XAI subcomponent in AI4CYBER.

management, static and dynamic application security testing or pen-testing analysis are used to protect the bank's infrastructure and for monitoring of security controls in the offered and used applications. CaixaBank's main focus is strengthening its defence mechanisms to be used against sophisticated network cyberattacks, as those attacks could result in the exposure of sensitive and personal data. Additionally, the bank aims at the detection and management of software vulnerabilities in a dynamic and evolving environment.

At CaixaBank, Identity and Access Management (IAM) is carried out with the suite of AIM/PAM (Access Identity Management/Privileged Access Management) tools designed to manage user access to different applications and services. From a security perspective, it is also important to note that bank applications, systems, and data are accessed not only by the bank's employees, but some of the environments are also accessible by third-party providers. Another important issue related to security is remote work, which impacts access control configuration to ensure employees can securely reach internal applications from their homes. The bank customizes different layers of IAM depending on the user type. As for log collection and monitoring, and incident response mechanisms, the bank utilizes SIEM and SOAR solutions. The Security Information and Event Management system provides monitoring and data analytics based on the logs collected from network assets (devices, applications), while the Security Orchestration, Automation, and Response (SOAR) system is used to automate response after detection and alerting of potential attacks.

For the purposes of the banking cybersecurity scenario, the proposed tools, i.e. AI4FIDS and TRUST4AI.XAI are sandboxed using the bank's Innovation Sandbox. The main source of the information is an interconnected SIEM solution feeding the isolated tools, while the abovementioned corporate SOAR solution is used to trigger tailored playbooks. The tools can also access internal development tools and repositories to analyse the code and development processes.

From the end-user viewpoint, CaixaBank adopts AI4CYBER tools for ensuring robust security in critical environments like the SWIFT client and the Financial Terminal. AI4FIDS enhanced by xAI capabilities, delivers AI services that can detect abnormal actions, identify impersonations of privileged users, and prevent intrusions and AI-driven attacks in real-time. In addition, the tools facilitate comprehensive monitoring of user behaviours and activities across different bank services. The xAI enhancement, provides insight into the model's output, making cybersecurity-related decisions transparent and understandable for the bank's security staff, saving time and cost of security operations.

EXPERIMENTS AND RESULTS

AI4FIDS – Initial Evaluation Results with Existing Cybersecurity Datasets

In order to demonstrate the efficacy and soundness of the initial version of AI4FIDS in conjunction with benchmark cybersecurity datasets, this section describes the system prerequisites and technical specifications of AI4FIDS. Multiple datasets were employed for evaluating the proposed implementation, yet in the context of this study, the distinct results for the aggregation techniques are provided based on the CSE CIC-IDS-2018 Dataset [31].

To that end, the initial results of the evaluation procedure of the respective containers, namely L-FIDS, O-FIDS, N-FIDS and V-FIDS, are presented. More specifically, T4FIDS is responsible for generating the federated models that will be utilized by the detection engines of the preceding containers. The testbed onto which the proposed implementation is trained on, is composed of three Federated Clients and one Federated Server. In the current study, preliminary results are provided for the N-FIDS container along with the CSE CIC-IDS-2018 Dataset [30]. In Table 1, the respective results from the evaluation procedure are shown when employing the network flow statistics calculated by the CICFlowMeter. As one may observe, a thorough inspection of a variety of aggregation methods is performed in the context of evaluating the initial version of AI4FIDS, where the N-FIDS container must detect attacks when the aforementioned features are considered. More notably, the detection engine, namely T4FIDS, is trained with the TCP/IP network flow statistics, where multiple cyberattacks are considered and five aggregation techniques are scrutinized. The attained results indicate that the finest performance was achieved from the FedProx technique. Additionally, in Fig. 3 the confusion matrix of the model that uses FedAvg with the TON IoT Dataset is illustrated.

Table 1. Performance comparison of different aggregation methods.

Aggregation	ACC	TPR	FPR	F1	AUC
FedAvg	84.97%	80.96%	1.13%	85.80%	98.60%
FedProx	86.73%	78.68%	1.01%	87.42%	98.17%
FedAdam	27.80%	35.66%	5.52%	28.11%	77.83%
FedAdagrad	85.66%	74.28%	1.10%	86.19%	97.86%
FedYogi	74.93%	71.01%	1.86%	77.22%	95.41%

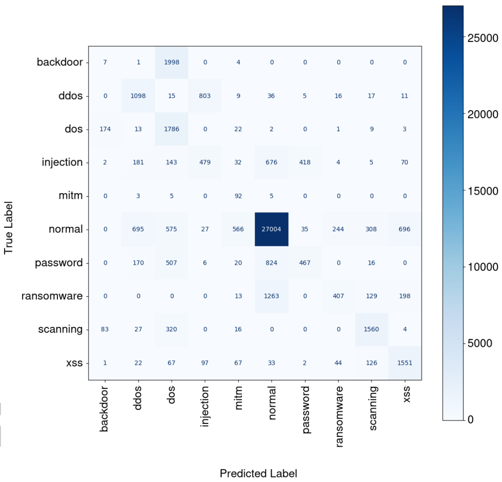

Figure 3. Confusion Matrix of the FL model that uses FedAvg with the TON IoT Dataset – Network Flow Statistics.

xAI Interpretation – Experiments and Results with Existing Cybersecurity Datasets

With the detection model established and evaluated, the experiment proceeded with attempting to gain insight into the decision-making process of the classifier. To this end, the xAI methods were employed. Following a scenario where a security operative needs to justify a decisive action to ban or not to ban a user based on the detection result, the xAI methods aim to provide reasoning as to why the samples in question were classified as an attack.

Fig. 4 demonstrates the application of the LIME technique within the TRUST4AI.XAI component of AI4CYBER. It shows how LIME decomposes a cybersecurity model's decision-making process for individual predictions, highlighting the contribution of each feature towards the predicted outcome of identifying network threats. Fig. 5 illustrates a decision tree that approximates the complex decision boundaries of the explained model, providing a simplified view of how various network statistics influence the classification of traffic. Fig. 6 showcases the SHAP explanations, each bar in the visualization represents the impact of an individual feature on the model's prediction, giving insight into which attributes are most influential for detecting cybersecurity threats. Fig. 7 shows ICE and PDP plots that analyze the impact of specific features on the predictions of the AI4FIDS cybersecurity model. These plots help in understanding the relationships between the feature values and the likelihood of an event being flagged as a security threat, across a range of values for the selected features.

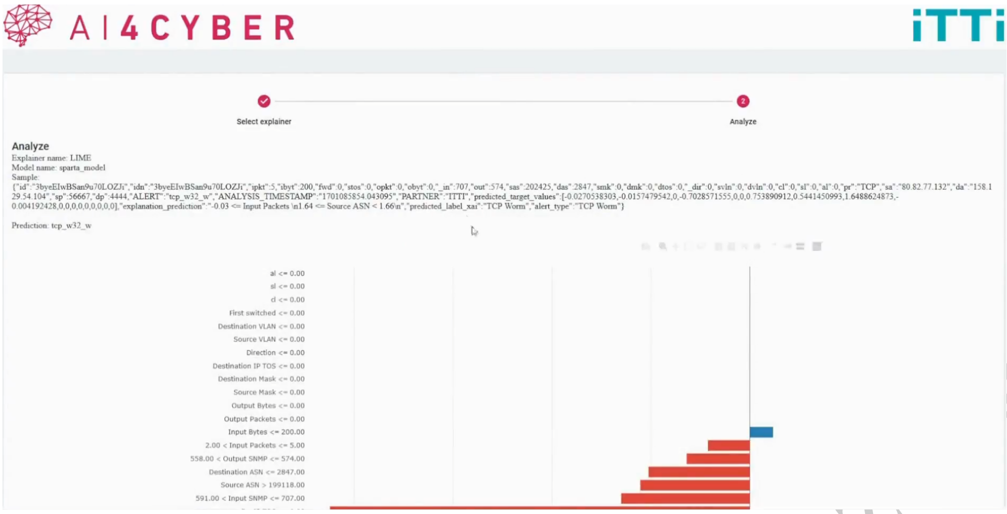

Figure 4. Examples of xAI methods: LIME explanations in the TRUST4AI.XAI component of AI4CYBER.

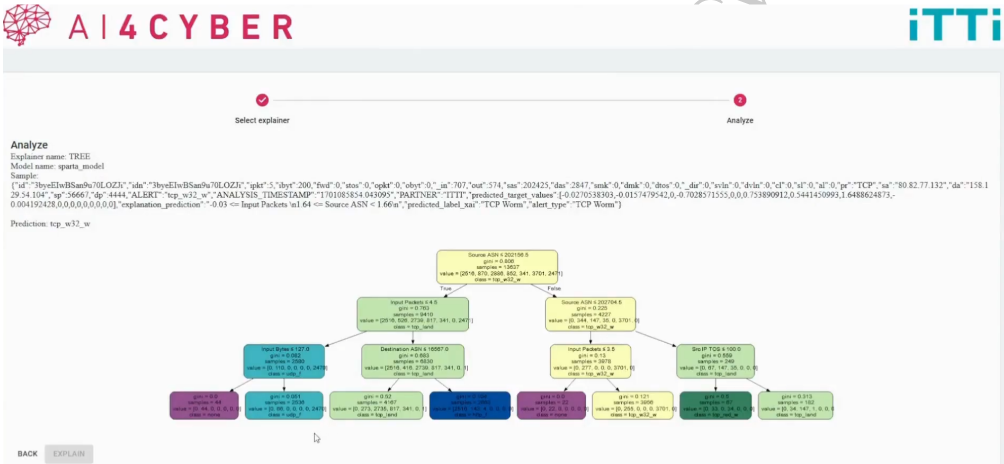

Figure 5. Examples of xAI methods: Surrogate Tree Aggregations in the TRUST4AI.XAI component of AI4CYBER.

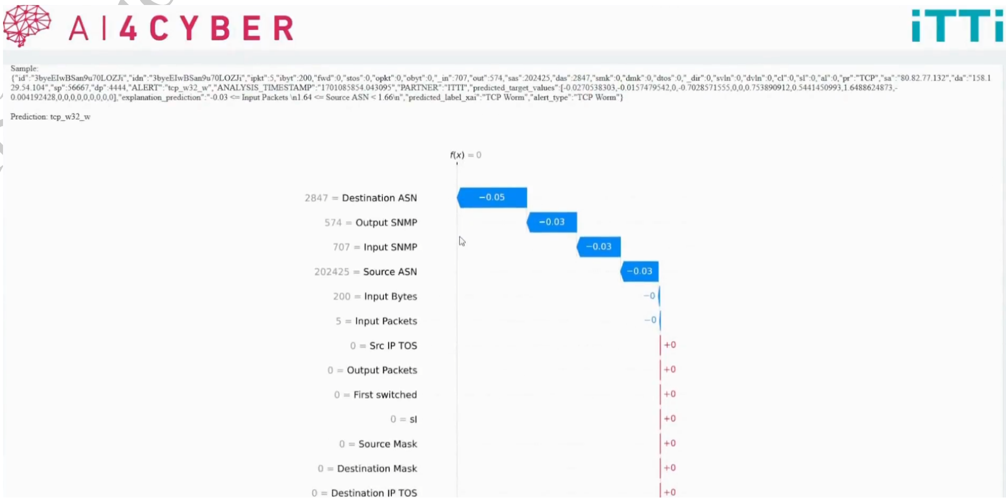

Figure 6. Examples of xAI methods: SHAP explanations in the TRUST4AI.XAI component of AI4CYBER.

Figure 7. Examples of xAI methods: ICE and PDP plots, explanations in the TRUST4AI.XAI component of AI4CYBER.

CONCLUSION

In this paper, AI4FIDS and TRUST4AI.XAI were introduced, two pivotal components of the AI4CYBER project aimed at enhancing cybersecurity in the financial sector through the utilization of advanced AI techniques and end-user-centric deployment in the CaixaBank pilot. AI4FIDS, empowered by Federated Learning, offers a privacy-preserving approach that enables robust intrusion detection across decentralized networks without compromising sensitive data. The TRUST4AI.XAI module provides critical explainability, ensuring that AI-driven decisions are transparent and interpretable to end-users. The sector-oriented experiments conducted with these systems within the financial pilot of CaixaBank demonstrate their efficacy in both detecting a range of cyberthreats and in aligning with requirements for transparency.

The implementation demonstrated the practical viability and effectiveness of combining Federated Learning with advanced explainability to secure financial infrastructures. The success of this integration not only confirms the potential of these technologies to improve cybersecurity practices but also sets a precedent for the development of privacy-preserving, interpretable AI models in the financial sector.

ACKNOWLEDGMENTS

This research is funded under the Horizon Europe AI4CYBER Project, which has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101070450.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[AI4CYBER] AI4CYBER. Ai4cyber. (Accessed on 07/22/2024).

[2] AL-Dosari, K., Fetais, N., and Kucukvar, M. (2024). Artificial intelligence and cyber defense system for banking industry: A qualitative study of ai applications and challenges. *Cybernetics and systems*, 55(2):302–330.

[3] Alazab, M., RM, S. P., Parimala, M., Maddikunta, P. K. R., Gadekallu, T. R., and Pham, Q.-V. (2021). Federated learning for cybersecurity: Concepts, challenges, and future directions. *IEEE Transactions on Industrial Informatics*, 18(5):3501–3509.

[4] Altmann, A., Toloşi, L., Sander, O., and Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347.

[5] Aouedi, O., Piamrat, K., Muller, G., and Singh, K. (2022). Fluids: Federated learning with semi-supervised approach for intrusion detection system. In *2022 IEEE 19th Annual Consumer Communications & Networking Conference (CCNC)*, pages 523–524. IEEE.

[6] Apley, D. W. and Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(4):1059–1086.

[7] Arisdakessian, S., Wahab, O. A., Mourad, A., Otrok, H., and Guizani, M. (2022). A survey on iot intrusion detection: Federated learning, game theory, social psychology, and explainable ai as future directions. *IEEE Internet of Things Journal*, 10(5):4059–4092.

[8] Campos, E. M., Saura, P. F., González-Vidal, A., Hernández-Ramos, J. L., Bernabe, J. B., Baldini, G., and Skarmeta, A. (2022). Evaluating federated learning for intrusion detection in internet of things: Review and challenges. *Computer Networks*, 203:108661.

[9] Choraś, M., Pawlicki, M., Puchalski, D., and Kozik, R. (2020). Machine learning–the results are not the only thing that matters! what about security, explainability and fairness? In *Computational Science–ICCS 2020: 20th International Conference, Amsterdam, The Netherlands, June 3–5, 2020, Proceedings, Part IV 20*, pages 615–628. Springer.

[Daly] Daly, R. A standout performance: Q&a with caixabank ceo gonzalo gortázar — global finance magazine. (Accessed on 09/06/2024).

[11] Dong, T., Qiu, H., Lu, J., Qiu, M., and Fan, C. (2021). Towards fast network intrusion detection based on efficiency-preserving federated learning. In *2021 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)*, pages 468–475. IEEE.

[12] Familoni, B. T. and Shoetan, P. O. (2024). Cybersecurity in the financial sector: a comparative analysis of the usa and nigeria. *Computer Science & IT Research Journal*, 5(4):850–877.

[13] Fan, Y., Li, Y., Zhan, M., Cui, H., and Zhang, Y. (2020). Iotdefender: A federated transfer learning intrusion detection framework for 5g iot. In *2020 IEEE 14th international conference on big data science and engineering (BigDataSE)*, pages 88–95. IEEE.

[14] Friha, O., Ferrag, M. A., Shu, L., Maglaras, L., Choo, K.-K. R., and Nafaa, M. (2022). Felids: Federated learning-based intrusion detection system for agricultural internet of things. *Journal of Parallel and Distributed Computing*, 165:17–31.

[15] Ghimire, B. and Rawat, D. B. (2022). Recent advances on federated learning for cybersecurity and cybersecurity for federated learning for internet of things. *IEEE Internet of Things Journal*, 9(11):8229–8249.

[16] Goldstein, A., Kapelner, A., Bleich, J., and Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *journal of Computational and Graphical Statistics*, 24(1):44–65.

[17] Greenwell, B. M., Boehmke, B. C., and McCarthy, A. J. (2018). A simple and effective model-based variable importance measure. *arXiv preprint arXiv:1805.04755*.

[18] Gurumoorthy, K. S., Dhurandhar, A., Cecchi, G., and Aggarwal, C. (2019). Efficient data representation by selecting prototypes with importance weights. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 260–269. IEEE.

[19] Lavaur, L., Pahl, M.-O., Busnel, Y., and Autrel, F. (2022). The evolution of federated learning-based intrusion detection and mitigation: a survey. *IEEE Transactions on Network and Service Management*, 19(3):2309–2332.

[20] Li, B., Wu, Y., Song, J., Lu, R., Li, T., and Zhao, L. (2020). Deepfed: Federated deep learning for intrusion detection in industrial cyber–physical systems. *IEEE Transactions on Industrial Informatics*, 17(8):5615–5624.

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

[22] Mirzaee, P. H., Shojafar, M., Pooranian, Z., Asefy, P., Cruickshank, H., and Tafazolli, R. (2021). Fids: A federated intrusion detection system for 5g smart metering network. In *2021 17th International Conference on Mobility, Sensing and Networking (MSN)*, pages 215–222. IEEE.

[23] Mothilal, R. K., Sharma, A., and Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 607–617.

[24] Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., Van Keulen, M., and Seifert, C. (2023). From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, 55(13s):1–42.

[25] Pawlicka, A., Choraś, M., and Pawlicki, M. (2020). Cyberspace threats: not only hackers and criminals. raising the awareness of selected unusual cyberspace actors-cybersecurity researchers' perspective. In *Proceedings of the 15th International Conference on Availability, Reliability and Security*, pages 1–11.

[26] Popoola, S. I., Gui, G., Adebisi, B., Hammoudeh, M., and Gacanin, H. (2021). Federated deep

learning for collaborative intrusion detection in heterogeneous networks. In *2021 IEEE 94th Vehicular Technology Conference (VTC2021-Fall)*, pages 1–6. IEEE.

[27] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). " 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*, pages 1135–1144.

[28] Ribeiro, M. T., Singh, S., and Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

[29] Schneble, W. and Thamilarasu, G. (2019). Attack detection using federated learning in medical cyber-physical systems. In *Proc. 28th Int. Conf. Comput. Commun. Netw.(ICCCN)*, volume 29, pages 1–8.

[30] Sharafaldin, I., Gharib, A., Lashkari, A. H., Ghorbani, A. A., et al. (2018a). Towards a reliable intrusion detection benchmark dataset. *Software Networking*, 2018(1):177–200.

[31] Sharafaldin, I., Lashkari, A. H., Ghorbani, A. A., et al. (2018b). Toward generating a new intrusion detection dataset and intrusion traffic characterization. *ICISSp*, 1:108–116.

[32] Uzougbo, N. S., Ikegwu, C. G., Adewusi, A. O., et al. (2024). Cybersecurity compliance in financial institutions: a comparative analysis of global standards and regulations. *International Journal of Science and Research Archive*, 12(1):533–548.

[33] Zhao, R., Wang, Y., Xue, Z., Ohtsuki, T., Adebisi, B., and Gui, G. (2022). Semisupervised federated-learning-based intrusion detection method for internet of things. *IEEE Internet of Things Journal*, 10(10):8645–8657.