Survey Research Paper on Explainable Federated Learning

Nazreen Shah
IIIT Delhi
PhD21122
nazreens@iiitd.ac.in

Sumedha Chugh
IIIT Delhi
PhD21123
sumedhac@iiitd.ac.in

***Abstract*—Today, as machine and deep learning algorithms are becoming data greedy, the need for training decentralized data while protecting data privacy has become a challenge; Federated Learning (FL) addresses this by introducing a distributed collaborative training of machine learning models placed at different clients. Various regulations and laws like GDPR are being established to enable individuals to understand how their data is being used and the right to the information that made the ML model produce a decision. Although FL is a favorable direction toward data privacy, it is responsible for further explaining the decisions made in an FL model. Explainable AI (XAI) is a way to achieve this and can offer a more responsible AI. Integrating XAI with FL can safeguard the interests of the users and can produce transparent and efficient privacy-preserving AI models. In this work, we survey Explainable Federated Learning techniques and find gaps in the explainability of FL methods. We discuss the causes for these gaps and how these can be overcome.**

***Index Terms*—Federated Learning, Explainable Federated Learning, Explainable AI, Collaborative Training, Responsible AI, Privacy Preservation**

I. Introduction

Federated learning (FL) is a machine learning (ML) paradigm that allows several devices/clients to communicate to a server and collaboratively build an ML model without disclosing its private data. This is accomplished by dividing the training process among the clients and having each client contribute its own local learning to the final ML model. However, in some cases, it may be necessary to understand how the federated learning model arrived at a particular decision. Explainable AI (XAI) improves understandability in federated learning (FL). Explainable federated learning (XFL) is a refined version of federated learning that focuses on making the model's decision-making process more transparent and understandable to users. It allows you to understand how the model made its decision by tracing and analyzing the contributions of each device/client or the central server to the final output. By allowing users to understand the decision-making process and identify any potential biases or inaccuracies, XFL can assist in strengthening the trustworthiness of the federated learning model. This is especially critical in areas like healthcare, banking, and security, where the model's judgments greatly impact people's lives. Model interpretation, visualization, and lineage tracking are examples of existing techniques. These strategies can help provide insights into the model's decision-making process and highlight any potential concerns. Overall, explainable federated learning is a significant advancement in the field of machine learning that can aid in increasing the openness and accountability of federated learning models, making them more reliable.

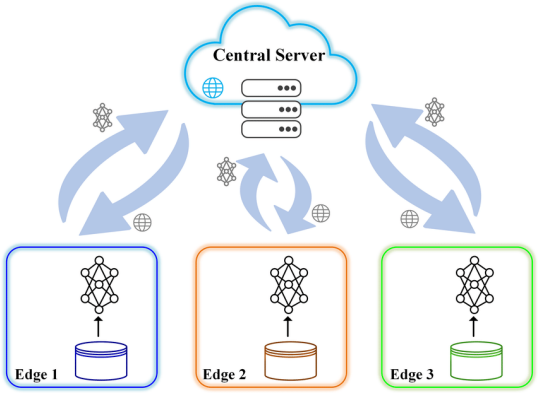

Fig. 1. Federated Learning (FL) [1]

II. Literature Review

As a new topic of interest, only limited techniques in the literature enable explainability in FL. A detailed explanation on the existing techniques is provided in this section. The XAI models used in FL scenarios can be mainly categorized into inherently explainable (ante-hoc) methods and post-hoc explainable methods. The literature review presented here is based on this categorization. In later sections, we point out the challenges of using inherently explainable methods and conclude by proposing a possible future research direction.

[2] provides an overview of the applications and challenges of FL-based approaches in the area of smart healthcare. The paper discusses the importance of federated learning and its potential benefits in healthcare, which includes improved privacy, scalability, and enhanced data quality. As a future direction, the authors propose integrating XAI into FL, as it only improves the trustworthiness of such an FL setup, especially in smart healthcare. The authors propose to use an inherently explainable XAI model in the FL set-up. Another paper [3] proposes an FL approach of fuzzy regression models that are both accurate and interpretable. Fuzzy regression models are inherently explainable and,

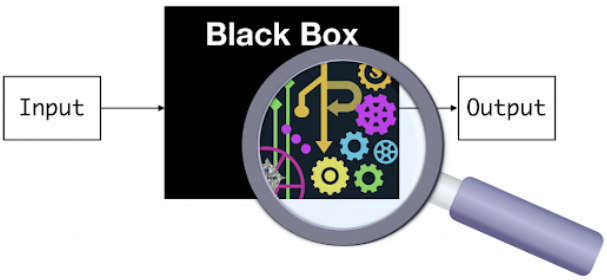

Fig. 2. Explainable AI

therefore, a promising explainable model that can be used in FL. The method uses a federated learning framework to train the fuzzy regression models on distributed data sources/clients. In particular, it uses Takagi-Sugeno- Kang Fuzzy Rule-based Systems, which promises high levels of interpretability. The integration of this inherently explainable XAI model is carried out on a vanilla FL strategy. In [4] discusses the potential of federated learning and explainable AI models in the area of 6G automated vehicle networking. Data privacy, security, and explainability are some problems connected with developing and deploying AI-based solutions in automated vehicle networking. The paper proposes a federated learning approach using explainable AI models to provide transparency and interpretability in decision-making. The suggested strategy addresses the obstacles to creating and deploying AI-based solutions in this domain while providing transparency and interpretability in autonomous vehicle decision-making. Particularly, the authors describe how they propose a one-shot communication-based FL using inherently explainable XAI models. Authors in [5] propose a new approach to bring explainability to vertical federated learning. It uses vertical data federation to enable sharing of specific data attributes while preserving data privacy. Because it combines vertical data federation and explainable machine learning models, the authors call it Explainable Vertical Federated Learning (EVFL). An explainability procedure called counterfactual explanation is being carried out in this paper. Another paper [6] proposes a new approach to federated learning that uses blockchain technology to enhance security and privacy in Internet-of-Things (IoT)-based social media networks, also known as social media 3.0. By introducing a blockchain based differentially private architecture, it offers an inherently explainable structure to the underlying FL strategy. In [7] proposes a federated learning method focusing on the industry life cycle, including data collection, model training, and model evaluation. In this paper, the proposed explainable feature uses a dashboard to visualize the entire process lifecycle, from data collection to model evaluation. Federated learning with a dashboard visualizing the entire process lifecycle is a promising development in the industrial area.

The work by Qiang et al. [8] aims to solve the problem of data sharing between multiple organizations for collaboration, as it can be a privacy issue. It explains the benefits of Federated Learning for reducing the cost of communication, scalability, and preserving privacy along with challenges faced like biases, lack of robustness proposing an idea to integrate Explainable AI techniques with FL settings. Another work that promotes collaboration-friendly Federated Learning is proposed by [9]. The authors present an algorithm called FederatedTrust which allows working on the project with careful sharing or no private data sharing. It does this by using a system that gives a trustworthiness score to each organization according to past behavior, generating a reputation system. One more advantage of the algorithm is that no single organization can access the complete model or the data, hence making it impossible to exploit user privacy. This method makes use of the FedEx Algorithm for FL and kernelSHAP for explainability and is widely used in applications based on Industry 4.0.

An extensively used application for FL is in Industrial Control Systems (ICS) for anomaly detection. Huong et al. [10] introduces a novel FL based framework for this application. They approach this problem by dividing it into three parts: feature extraction, collaborative training and explainability. For feature extraction they make use of Deep Neural Networks; for collaborative training Federated Learning based FedeX algorithm is utilized and finally post-hoc explainability using SHAP is attained. Another area where collaborative training is highly advantageous is taxi travel time prediction. This problem aims to calculate the time taken by taxis from one point to another. [11] works on a similar approach as [10], enabling multiple companies to train a travel time prediction model in alliance without the need of sharing private, leading to more secure and effective predictions. They use FedAvg for training and Integrated Gradients-based ad-hoc method for explainability. Both of these methods on respective real-world dataset show high interpretability and superior accuracy, outperforming traditional methods. Another safety critical domain that can benefit highly from Federated Learning, like training and explainability is healthcare. [12] leverages XFL for electrocardiogram (ECG) monitoring, enabling multiple hospitals to cooperate for training using transfer learning, leading to highly accurate diagnosis while preserving patient privacy. For explainability, they use GradCAM as well as attention based methods. This work is domain-specific, but can be extended to various medical domains. One such domain adaptive FL based approach for medical diagnosis is suggested by [13]. This method is easily interpretable, using feature relevance analysis and Visual explanations. For domain adaptation, authors use transfer learning that allows this work to be adapted in various healthcare based domains and Horizontal FL allows multiple institutes to train simultaneously giving a safer, more accurate and explainable model that can be used irrespective of the healthcare domain. A detailed explanation of these methods can be found in the following section for a clear understanding of these works.

III. Methods

Federated learning starts by initializing the models at the server and clients. The clients conduct local training with the available data and send the updated weights to the server, where it aggregates the weights to produce the global model. The global model is again sent back to the clients for further updates. Communication is usually periodic and limited. The communication of the updated model weights is always carried out in a way that accomplishes privacy. Broadly, FL is categorized as horizontal FL where the data has a similar feature set across clients, and vertical FL where the data has a different feature set across clients. One of the most famous algorithms in FL is FedAvg [14]. It is known to be the simplest among the existing techniques, and yet it achieves the goal of FL.
Federated Learning is being utilized in many important areas, including smart healthcare. In these domains, the trustworthiness of the ML models is a much-needed aspect. In 2018 European Union (EU) passed General Data Protection and Regulation (GDPR) law which gives every user the right to know the explanations for the decisions made by Machine Learning systems. This law is not very easy to follow, as the Machine Learning models are becoming increasingly complex. Due to the increase in this complexity, it is not possible for even engineers and data scientists to explain which features the algorithm used to make predictions. Some popular works have shown how highly accurate Machine and Deep learning models made use of completely irrelevant features like background or noise to make predictions. This can be dangerous for domains where health, wealth or time are at stake. Explainable AI methods are being employed now with ML and DL methods to know which features contributed to the decision-making, hence reducing the risk factor. Keeping the safety critical domain of Federated Learning and the reliability aspect of explainable AI, we studied various mechanisms on Explainable Federated Learning (XFL), leading to a more responsible AI. Based on the literature survey, we came across the two most important methods to summarize Explainable Federated Learning: Inherently Explainable and Post-hoc Methods. For inherently explainable methods, the ML model placed at the clients is an interpretable one whose output humans can easily understand. Examples of such models are blockchains, fuzzy rule-based methods, dashboard visualizations, decision trees, linear regression, logistic regression, etc. These simple ML techniques are aggregated with Federated Learning modules to give results that don't require any more processing and generate results that are human-understandable and hence explainable. The functioning of this method is shown in Fig.3.

These simple models are human-understandable but might not give good results; deep learning models, such as neural networks, VAEs, etc., give much better results but are not illustrated. In Federated Learning settings where one can't risk accuracy, using inherently explainable methods is not recommended. To get explainability in such cases, posthoc XAI methods are used after the FL model has been deployed. Fig.4 shows the working of FL methods with Post-hoc methods. Methods used with Federated Learning are LIME, SHAP, GradCAM, GradCAM++, Feature importance map etc.

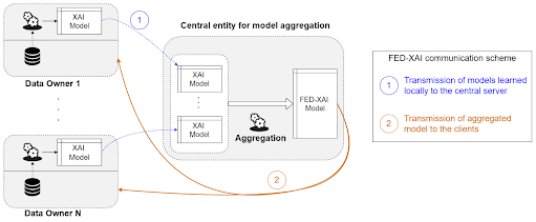

Fig. 3. Inherently Explainable FL [4]

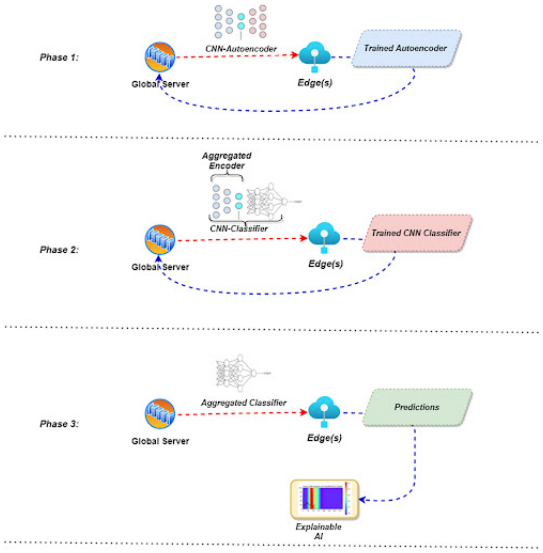

Fig. 4. Post-hoc FL [9]

Some commonly used methods from existing works are listed in table I. The works have been classified based on the application, FL, and XAI features. It is clear from the table that XFL is robust enough and is being utilized in various safety-critical and domain-specific applications such as Smart healthcare, 6-G automation, industry 4.0, travel time prediction, etc.

IV. Discussion

FL has been very advantageous in the areas of healthcare and finance, but the lack of explainability of deep learning models used in FL has reduced transparency and hence the usability and trustworthiness of these in critical domains, where the outputs of these methods significantly affect decision-making. As discussed in the methods section, inherently explainable models can be a reasonable replacement for Deep learning models for increased explainability. But from section 2 and section 3, we found out that FL-based methods that make use of inherent explainability are significantly less. However, using these ante-hoc methods can be challenging in federated settings. We devised the cause for this as the absence of a global objective that is differentiable. The basic step of an FL algorithm is to minimize this global objective, which is an aggregation of local objectives. As inherent XAI

TABLE I
FL AND XAI FEATURES USED IN THE EXISTING WORK

Work	Application	FL feature	XAI feature
[2]	Smart healthcare	For privacy and security	Inherently explainable
[3]	Regression problems	Vanilla FL	Fuzzy Rule-based Systems
[4]	6G-Automated vehicle networks	One shot communication FL	Inherently XAI
[5]	Data oriented AI systems	Vertical FL	Counterfactual explanation
[6]	Social media 3.0(IoT based)	Blockchain based, differentially private	Inherently explainable due to blockchains
[7]	Industry settings	Industry FL (IoT)	Dashboard visualization
[9]	Trustworthy FL	FederatedScope	Feature Importance MAP
[10]	Industrial Control System	FedeX	SHAP
[15]	COVID-19 Detection	FedMoCo	GradCAM++
[13]	Personal Health Care	Horizontal FL	Feature relevance analysis and Visual explanation
[12]	ECG Based Healthcare	Federated Transfer Learning: FedMod	GradCAM
[11]	Taxi Travel Time Prediction	FedAverage	Integrated Gradients

methods like decision tree, blockchains etc. can't offer a global objective, we need to come up with novel suitable aggregation strategies. Another challenge is the if-else modeling of rule-based methods that makes it impossible to integrate with FL using basic FL aggregation. Integrating rule-based methods is largely different from integrating Neural Networks-based models. One will have to use methods like rule extraction to come up with an aggregated version of the rules proposed by the clients. These methods are also very simple and can not capture the underlying complexities of Federated Learning Algorithms. To overcome these challenges, we propose using a type of hybrid learning method that combines both rule-based interpretable ways with deep learning models, which ensures a more trustable method with high accuracy.

V. CONCLUSION

At present, the AI community is increasingly interested in making AI more responsible and trustworthy. Explainable Federated Learning is unquestionably a big leap in this direction. By merging the explainability and interpretability features of XAI with the privacy and security features of FL, the proposed idea is enabling AI to become more transparent and unambiguous. As a part of this literature survey, we have presented a detailed analysis of how XAI and FL are combined in areas of healthcare, Industry, Automated vehicles, etc. After a thorough survey of the existing literature, it is clear that there is a major gap in the application of inherently explainable methods in FL, and we have formulated the causes for this gap. There is a need to devise methods that can enable such transparent models in the FL architecture. As a future scope of research, these challenges in integrating XAI and FL need to be tackled.

REFERENCES

[1] Nastaran Gholizadeh and Petr Musilek, "Distributed learning applications in power systems: A review of methods, gaps, and challenges," *Energies*, vol. 14, no. 12, pp. 3654, 2021.

[2] Anichur Rahman, Md Sazzad Hossain, Ghulam Muhammad, Dipanjali Kundu, Tanoy Debnath, Muaz Rahman, Md Saikat Islam Khan, Prayag Tiwari, and Shahab S Band, "Federated learning-based ai approaches in smart healthcare: concepts, taxonomies, challenges and open issues," *Cluster Computing*, pp. 1–41, 2022.

[3] José Luis Corcuera Bárcena, Pietro Ducange, Alessio Ercolani, Francesco Marcelloni, and Alessandro Renda, "An approach to federated learning of explainable fuzzy regression models," in *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2022, pp. 1–8.

[4] Alessandro Renda, Pietro Ducange, Francesco Marcelloni, Dario Sabella, Miltiadis C Filippou, Giovanni Nardini, Giovanni Stea, Antonio Virdis, Davide Micheli, Damiano Rapone, et al., "Federated learning of explainable ai models in 6g systems: Towards secure and automated vehicle networking," *Information*, vol. 13, no. 8, pp. 395, 2022.

[5] Peng Chen, Xin Du, Zhihui Lu, Jie Wu, and Patrick CK Hung, "Evfl: An explainable vertical federated learning for data-oriented artificial intelligence systems," *Journal of Systems Architecture*, vol. 126, pp. 102474, 2022.

[6] Sara Salim, Benjamin Turnbull, and Nour Moustafa, "A blockchain-enabled explainable federated learning for securing internet-of-things-based social media 3.0 networks," *IEEE Transactions on Computational Social Systems*, 2021.

[7] Michael Ungersböck, Thomas Hiessl, Daniel Schall, and Florian Michahelles, "Explainable federated learning: A lifecycle dashboard for industrial settings," *IEEE Pervasive Computing*, 2023.

[8] Qiang Yang, "Toward responsible ai: An overview of federated learning for user-centered privacy-preserving computing," *ACM Trans. Interact. Intell. Syst.*, vol. 11, no. 3–4, oct 2021.

[9] Pedro Miguel Sánchez Sánchez, Alberto Huertas Celdrán, Ning Xie, Gérôme Bovet, Gregorio Martínez Pérez, and Burkhard Stiller, "Federatedtrust: A solution for trustworthy federated learning," 2023.

[10] Truong Thu Huong, Ta Phuong Bac, Kieu Ngan Ha, Nguyen Viet Hoang, Nguyen Xuan Hoang, Nguyen Tai Hung, and Kim Phuc Tran, "Federated learning-based explainable anomaly detection for industrial control systems," *IEEE Access*, vol. 10, pp. 53854–53872, 2022.

[11] Jelena Fiosina, "Explainable federated learning for taxi travel time prediction," in *International Conference on Vehicle Technology and Intelligent Transport Systems*, 2021.

[12] Ali Raza, Kim Phuc Tran, Ludovic Koehl, and Shujun Li, "Designing ecg monitoring healthcare system with federated transfer learning and explainable ai," *Knowledge-Based Systems*, vol. 236, pp. 107763, 2022.

[13] Ahmad Chaddad, Qizong Lu, Jiali Li, Yousef Katib, Reem Kateb, Camel Tanougast, Ahmed Bouridane, and Ahmed Abdulkadir, "Explainable, domain-adaptive, and federated artificial intelligence in medicine," *IEEE/CAA Journal of Automatica Sinica*, vol. 10, pp. 859–876, 04 2023.

[14] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[15] Nanqing Dong and Irina Voiculescu, "Federated contrastive learning for decentralized unlabeled medical images," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*. Springer, 2021, pp. 378–387.