JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

INTEGRATING EXPLAINABILITY AND SECURITY IN FEDERATED LEARNING FOR EDGE COMPUTING: CHALLENGES AND SOLUTIONS

[1]Dr. Supriya Shree, Assistant Professor, Department of Computer Science, St. Xavier's College of Management & Technology, Patna.

[2]Ms. Vandana Verma, Assistant Professor, Department of Computer Science, St. Xavier's College of Management & Technology, Patna.

ABSTRACT

The growing adoption of Federated Learning (FL) in edge computing environments has transformed how intelligent systems are trained by ensuring privacy-preserving and bandwidth-efficient model development. However, the inherently decentralized and non-IID nature of FL presents unique challenges in model transparency and interpretability, necessitating the integration of Explainable Artificial Intelligence (XAI). This paper explores how existing XAI techniques can be effectively adapted to the federated learning paradigm, particularly focusing on privacy preservation and communication efficiency. It analyzes the limitations of traditional explainability methods in distributed settings and proposes novel adaptations, including the use of local surrogate models, federated explainability aggregation, and differentially private explanations. Furthermore, it examines advanced cryptographic techniques, such as Secure Multiparty Computation (SMPC) and homomorphic encryption, to enable secure explanation sharing without compromising sensitive data. Through a comprehensive study of methods like SHAP, LIME, Grad-CAM, and counterfactual explanations, the research provides a systematic framework for building trustworthy and interpretable federated learning systems. This work offers valuable insights for developing next-generation AI applications that are both transparent and privacy-conscious, particularly in sensitive domains such as healthcare, finance, and smart cities.

Keywords: Secure Aggregation, Adversarial Attacks, Model Poisoning, Data Poisoning, Differential Privacy.

I. INTRODUCTION

Federated Learning (FL) is an innovative machine learning paradigm that enables multiple decentralized devices or nodes to collaboratively train a global model without sharing their raw data. This approach is particularly well-suited for edge computing environments, where data is generated at the edge of the network, such as in mobile

devices, IoT sensors, and distributed computing nodes. Unlike traditional centralized machine learning models that require data aggregation on a central server, FL ensures that training occurs locally on edge devices, preserving data privacy and reducing network congestion. However, while FL offers significant advantages in terms of efficiency and privacy, it also introduces a range of security vulnerabilities that can compromise the integrity, confidentiality, and availability of the learning process. Ensuring security in FL within edge computing is therefore critical to unlocking its full potential for real-world applications.

The rise of edge computing has been driven by the increasing need for low-latency, high-efficiency data processing. In traditional cloud-based computing models, data from end-user devices is transmitted to centralized servers, where computation and analysis take place. This model is often inefficient due to high bandwidth consumption, latency issues, and security concerns associated with data transmission. Edge computing addresses these issues by bringing computation closer to the data source, enabling faster processing and reduced reliance on cloud infrastructure. Federated Learning integrates seamlessly with edge computing by enabling distributed devices to learn from local data while maintaining autonomy over their datasets. This integration enhances real-time processing capabilities, making FL particularly valuable for applications in healthcare, smart cities, autonomous vehicles, and industrial IoT. However, the decentralized nature of FL also introduces new attack vectors that adversaries can exploit to manipulate or extract sensitive information from models, making security a primary concern.

One of the primary security challenges in FL is data poisoning, where malicious participants inject corrupt data into the training process to degrade the global model's performance. Since FL relies on updates from multiple clients, an adversary can intentionally provide misleading gradients that skew the learning process, leading to biased or inaccurate model predictions. Similarly, model poisoning attacks occur when an attacker alters model updates to introduce vulnerabilities, such as backdoors, that can be exploited later. These types of attacks are particularly dangerous in security-sensitive applications like fraud detection, medical diagnosis, and autonomous driving, where model integrity is crucial. In addition to poisoning attacks, FL is also susceptible to inference attacks, where adversaries attempt to reconstruct private training data by analyzing model updates. This poses significant privacy risks, as sensitive user information can be extracted without directly accessing raw data.

Another major concern in secure federated learning is ensuring confidentiality and integrity during model aggregation. In a typical FL setting, local updates are sent to a central server for aggregation, which creates opportunities for adversarial interference. If an attacker gains access to the aggregation process, they can manipulate updates, alter model weights, or introduce biases that impact decision-making. Moreover, eavesdropping attacks can occur during the communication of model updates, allowing unauthorized parties to intercept and analyze transmitted data. Since edge devices often operate in dynamic and heterogeneous environments with varying levels of security, the risk of compromised clients participating in FL is high. Traditional cryptographic methods, such as encryption, can provide a layer of protection, but they introduce computational overhead that may not be feasible for resource-constrained edge devices. Therefore, lightweight

and efficient security mechanisms are required to ensure the confidentiality and integrity of model updates while minimizing computational costs.

To address these challenges, researchers have explored various security-enhancing techniques for FL in edge computing. One of the most widely adopted approaches is differential privacy, which adds carefully calibrated noise to model updates to prevent adversaries from extracting sensitive information. Differential privacy ensures that individual contributions remain indistinguishable, thereby enhancing user privacy while maintaining model accuracy. Another promising technique is homomorphic encryption, which enables computations to be performed on encrypted data, ensuring that model updates remain confidential even during aggregation. However, the computational complexity of homomorphic encryption makes it challenging to implement in large-scale FL networks. Secure Multi-Party Computation (SMPC) is another approach that allows multiple participants to collaboratively compute a function over their inputs without revealing their individual data. SMPC provides strong security guarantees but often comes with increased communication overhead, which can be a limiting factor in edge environments.

Blockchain technology has also emerged as a potential solution for securing FL in edge computing. By leveraging decentralized consensus mechanisms, blockchain can ensure that model updates are transparently recorded and verified, preventing malicious modifications. Smart contracts can be used to enforce security policies, validate client contributions, and detect anomalies in model updates. Additionally, blockchain-based FL can mitigate the risk of a single point of failure by distributing trust across multiple nodes, making the system more resilient to attacks. However, integrating blockchain with FL introduces scalability concerns, as blockchain networks require significant computational resources and storage capacity. Researchers are actively exploring lightweight blockchain frameworks that can be optimized for edge computing environments.

Another crucial aspect of securing FL in edge computing is developing robust defense mechanisms against adversarial attacks. Techniques such as Byzantine-resilient aggregation aim to filter out malicious updates by identifying and discarding anomalous contributions. Federated Averaging (FedAvg), a commonly used aggregation algorithm, can be enhanced with outlier detection methods to prevent adversarial interference. Additionally, trust-based client selection mechanisms can be implemented to ensure that only reliable and authenticated participants contribute to model updates. These approaches help maintain the integrity of FL while minimizing the impact of malicious actors.

Despite the advancements in secure federated learning, several challenges remain. One of the biggest hurdles is scalability, as securing FL in large-scale edge networks requires efficient and adaptive security mechanisms that can handle a high volume of model updates. Moreover, energy constraints in edge devices pose a challenge for implementing complex cryptographic techniques, necessitating the development of lightweight security solutions. The dynamic nature of edge environments, where devices frequently join and leave the network, further complicates security management. Ensuring that new participants are authenticated and trustworthy while preventing malicious entities from infiltrating the system requires robust access control mechanisms.

Additionally, achieving a balance between security and model performance remains a key concern, as overly restrictive security measures can degrade learning efficiency.

Looking ahead, the future of secure FL in edge computing will likely involve a combination of advanced cryptographic techniques, decentralized trust mechanisms, and adaptive security frameworks. Researchers are exploring novel methods such as secure enclave computing, which leverages trusted execution environments (TEEs) to isolate and protect sensitive computations. AI-driven anomaly detection can also enhance FL security by identifying suspicious behavior in model updates in real time. Moreover, federated reinforcement learning, which extends FL to reinforcement learning settings, introduces new security considerations that require further investigation. As FL continues to evolve, addressing security challenges will be essential for its successful deployment in real-world applications.

In secure federated learning in edge computing is a critical research area that addresses the need for privacy-preserving, efficient, and decentralized AI solutions. While FL provides significant advantages by enabling collaborative learning without sharing raw data, its security vulnerabilities pose serious risks to model integrity, privacy, and reliability. Adversarial attacks, inference threats, and communication security challenges necessitate the development of robust defense mechanisms, including differential privacy, homomorphic encryption, blockchain-based security, and Byzantine-resilient aggregation. However, challenges related to scalability, computational efficiency, and dynamic participation must be addressed to ensure the widespread adoption of secure FL in edge computing. As research in this domain progresses, interdisciplinary approaches combining machine learning, cryptography, and distributed systems will play a crucial role in strengthening FL security. By advancing security frameworks and adaptive defense mechanisms, federated learning can become a trustworthy and scalable solution for AI-driven edge computing applications across diverse industries.

Research Objectives

1. To Investigate how explainability techniques can be applied or adapted to federated settings.
2. To analyze XAI methods that are privacy-preserving and communication-efficient.

II. REVIEW OF RELATED STUDIES

Mishra, Sambit et al., (2023) In order to improve federated learning (FL), this research investigates several ways to use edge computing. We look at three methods: Edge-Fed, cluster federated learning, and hybrid federated learning at edge devices. By moving computations to edge servers, the Edge-Fed method solves the computational and communication problems that mobile devices encounter in FL. It presents a network design that allows for local aggregations while decreasing the frequency of global communication by use of Internet of Things (IoT) devices, an edge server, and a central cloud server. Advantages of Edge-Fed include lower computing costs, quicker training, and lower bandwidth needs. One way to improve FL in MAEC systems is by using hybrid federated learning at edge devices. Cluster federated learning improves FL performance by introducing a hierarchical aggregation mechanism based on clusters. Smart cities, healthcare, cybersecurity, autonomous cars, natural language processing, and vehicular networks are just a few of the sectors that these concepts may be used

to in this study. Integrating EC with FL is an exciting new approach that is quickly gaining traction in a wide variety of domains. EC improves FL even more by bringing cloud computing services in close proximity to data sources. When it comes to group projects, there may be advantages to combining FL with EC.

Duan, Qiang et al., (2023) Ubiquitous Intelligence (UI) is critical for 6G networks to fully harness the massive amounts of data produced by a high number of user devices in order to provide intelligent services. Collaborative Machine Learning (ML) relies on the data and computing resources provided by a vast number of network devices, which are essential to the development of user interfaces (UIs). A novel machine learning technique called Federated Learning (FL) lets data owners work together to train models without disclosing sensitive information. This means that user devices may contribute data to the development of user interfaces. In order to facilitate FL, edge computing brings cloud-like capabilities to the periphery of the network, allowing nodes in the network to volunteer their processing resources. So, it's possible that the 6G network's development of ubiquitous intelligence might be significantly aided by a mix of FL and edge computing. Taking a bird's-eye view of both FL and edge computing, this article provides a thorough overview of the latest innovations in both areas' ability to work together. For this study, we considered two different vantage points: that of a FL framework operating in an edge computing setting (FL in Edge) and that of an edge computing system that hosts FL (Edge for FL). We begin by outlining the primary obstacles to FL in edge computing from the FL in Edge vantage point, and then we take stock of the most prominent technological solutions to these problems. First, we take a look at the major needs for edge computing to facilitate FL from an Edge for FL standpoint. Then, we take a look at the new developments in edge computing that may be used to fulfill those needs. In order to spark interest in this new and fascinating multidisciplinary area, we go on to address outstanding difficulties and propose some potential future research areas regarding the combination of FL and edge computing.

Brecko, Alexander et al., (2022) As a result of recent technological developments, edge devices may now learn basic models that can be used in practice, bringing AI and ML to the periphery of the network. The goal of the distributed machine learning method known as federated learning (FL) is to build a global model using data acquired from several dispersed edge clients. While FL techniques have many benefits, such as scalability and data privacy, they also come with some dangers and downsides, particularly when it comes to computational complexity and heterogeneous devices. The processing power, network speed, or operating system of Internet of Things (IoT) devices could be lower than that of traditional computer systems. Focusing on edge devices with low computing capabilities, this study offers a summary of the methodologies employed in FL. Also included in this study are common FL frameworks that facilitate client-server communication. Contributions and current developments in the field are among the many subjects covered under this framework. Included in this are the fundamentals of system architectural models and designs, potential practical applications, security and privacy, and the administration of resources. We go over some of the problems with edge devices' computational demands, including hardware heterogeneity, communication overload, and device resource limitations.

Zhang, Zhuangzhuang et al., (2022) In the present day, a large number of suppliers of edge computing services anticipate improving their models via the use of data and processing capacity of edge nodes without transferring

any data. Distributed edge nodes may work together to build global models using federated learning, even if they don't have to share training data. Using current privacy-preserving federated learning in this context still encounters three problems, unfortunately: 3) Edge nodes have limited processing capacity and may drop out often; 4) it cannot provide Byzantine robustness while keeping data privacy; 5) it usually uses sophisticated cryptographic techniques, which causes significant training cost. To sum up, edge computing situations are not suitable for the privacy-preserving federated learning. Hence, we lay out LSFL, a privacy-preserving and Byzantine-Robust federated learning method, to be both safe and lightweight. To be more precise, we develop the Lightweight Two-Server safe Aggregation protocol that makes use of two servers to provide safe Byzantine robustness and combine models. This method safeguards sensitive information and prevents Byzantine nodes from interfering with model aggregation. The experiment findings demonstrate that LSFL achieves the design objectives of fidelity, security, and efficiency while maintaining model accuracy when compared to the popular FedAvg method, and we conduct the evaluation and implementation in a LAN setting.

Emmanni, Phani Sekhar. (2021). In order to tackle complicated cybersecurity issues, the study delves into how federated learning may be integrated with cloud and edge computing frameworks. Improving cybersecurity may be achieved via the use of federated learning, which allows for collaborative model training across decentralized devices without data sharing. In the context of the expanding Internet of Things (IoT) ecosystem, this research explores the possibility of federated learning to enhance real-time threat detection capabilities and data analysis that preserves privacy. This study uses a comprehensive literature review and theoretical analysis to explore the theoretical underpinnings of federated learning and its potential uses in cybersecurity settings. In order to improve data privacy and decrease latency in threat detection, it assesses the pros and cons of federated learning. In addition, the study uses qualitative analysis to evaluate the technical and security difficulties of federated learning implementation, such as communication cost, model aggregation complexity, and susceptibility to poisoning. According to the research, federated learning greatly increases privacy-preserving data analysis and real-time threat detection capabilities. This is because it allows collaborative learning while keeping data localized. On the other hand, it highlights important difficulties in federated learning strategy deployment, including communication cost, model aggregation complexity, and the possibility of model poisoning. In order to make full use of federated learning in cybersecurity, the study emphasizes the requirement of strong methods to handle these issues.

Liu, Gaoyang et al., (2021) The capacity of edge computing to provide cloud computing services and utilities to the periphery of networks while maintaining low communication costs and reaction times has lately garnered considerable attention. In most cases, mobile users are required to send their raw data to a centralized data server in order for edge computing to proceed. Nevertheless, this data often includes private information about mobile users, such as their sexual orientation, political beliefs, health, and past usage of services, that the users would prefer not to have publicized. Due to the increased number of devices that may access user data during transmission, the danger of data leakage increases. Here, we take a stab at preventing user privacy leaks by recommending that edge devices and end users retain data on their local storages. With this goal in mind, we present P2FEC, a privacy-preserving framework that combines federated learning with edge computing. It can

build a single deep learning model that spans several users or devices without transferring data to a central server. For the purpose of analyzing edge computing's privacy, we also use membership inference attacks. The results demonstrate that our framework-built model outperforms the typical edge computing-trained model in terms of prediction accuracy while providing more stringent data privacy protections.

Ye, Yunfan et al., (2020) An developing method to train a deep learning model using decentralized data, federated learning (FL) has garnered significant interest with the growth of mobile internet technologies. Unfortunately, computing capabilities are sometimes constrained by hardware limitations, and modern mobile devices frequently have access to rich yet privacy-sensitive data. Prior research using the federated averaging (FedAvg) method shown that global communication was slowed down because mobile devices had to do a large number of computations. Our proposal, edge federated learning (EdgeFed), is based on edge computing and divides up the task of updating the local model, which is meant to be done independently by mobile devices. For better learning efficiency and lower global communication frequency, the edge server aggregates the outputs of mobile devices. In many bandwidth conditions, our suggested EdgeFed has been shown to have benefits in empirical trials. In particular, in comparison to FedAvg, the computational cost of mobile devices and the global connection expenditure may be concurrently decreased by shifting some of the computations to the edge server.

III. FEDERATED LEARNING IN EDGE COMPUTING

Definition: Federated Learning (FL) is a decentralized machine learning technique where multiple edge devices work together to train a global model without sharing their raw data. Unlike traditional machine learning approaches, which require centralized data collection, FL ensures that data remains on local devices, improving privacy and reducing the need for large-scale data transfers. This approach is particularly useful in scenarios where data privacy is a concern, such as healthcare, finance, and personalized applications. By transmitting only model updates rather than entire datasets, FL significantly reduces bandwidth consumption, making it more efficient for distributed computing environments.

Role of Edge Computing: Edge computing refers to the practice of processing data closer to its source—such as IoT devices, smartphones, and sensors—rather than relying on a centralized cloud infrastructure. This proximity to data sources reduces latency, enhances real-time processing, and minimizes network congestion. Given that FL operates on distributed edge devices, edge computing provides an ideal platform for deploying FL models. The integration of FL with edge computing allows intelligent systems to make quick, data-driven decisions without excessive dependence on cloud services. This is particularly beneficial in applications requiring low-latency responses, such as autonomous vehicles, industrial automation, and real-time healthcare monitoring.

Advantages of FL in Edge Computing:

- **Privacy Preservation**: One of the most significant benefits of FL in edge computing is its ability to maintain user privacy. Since raw data never leaves the local device, the risk of data breaches, unauthorized access, and privacy violations is substantially reduced. This is especially crucial in sensitive applications such as medical diagnosis and financial transactions, where user confidentiality must be maintained.

- **Reduced Latency**: Because data processing occurs locally on edge devices, FL minimizes the time required for model updates and predictions. Traditional cloud-based machine learning solutions often experience delays due to data transmission and server processing times. With FL, model training and inference happen closer to the source, enabling faster decision-making in real-time applications such as autonomous driving and smart surveillance.

- **Bandwidth Efficiency**: In conventional centralized machine learning, raw data must be continuously uploaded to a central server for processing, leading to high network bandwidth consumption. FL, however, only transmits model updates (such as gradient updates or weight changes), significantly reducing network congestion and improving efficiency, especially in bandwidth-constrained environments.

- **Personalization**: FL enables models to learn from user-specific data on individual edge devices, allowing for highly personalized experiences without sacrificing privacy. This is particularly useful in applications like personalized healthcare, recommendation systems, and adaptive learning platforms, where models need to be tailored to the unique behavior and preferences of users.

Security Challenges in FL at the Edge:

Despite its advantages, FL in edge computing is vulnerable to several security threats, which must be addressed to ensure reliable and secure model training.

- **Data Poisoning Attacks**: In a data poisoning attack, malicious devices inject manipulated or incorrect data into the training process to distort the model's performance. This can lead to biased or incorrect predictions, which may have severe consequences in critical applications like healthcare and autonomous driving.

- **Model Poisoning**: Attackers can alter the model updates being transmitted from edge devices, introducing vulnerabilities that compromise the model's integrity. By manipulating gradient updates or injecting backdoor behaviors, attackers can make the model behave maliciously while appearing normal.

- **Inference Attacks**: Even though FL does not transmit raw data, attackers can still analyze the shared model updates to infer sensitive information about individual users. By examining the gradients or parameters of the model, adversaries may reconstruct data patterns or extract private information.

- **Communication Threats**: FL relies on network communication to exchange model updates between edge devices and the central server. This opens the system to potential threats such as eavesdropping, where attackers intercept transmitted updates, and man-in-the-middle attacks, where adversaries alter data before it reaches its destination. These threats can compromise data integrity and confidentiality.

Security Solutions for FL in Edge Computing:

To mitigate the security risks associated with FL, several advanced techniques have been developed to enhance privacy, confidentiality, and data integrity.

- **Differential Privacy**: This technique adds carefully calibrated noise to model updates before transmission, preventing attackers from extracting specific information about individual data points. Differential privacy ensures that even if an adversary gains access to model updates, they cannot accurately reconstruct the original data.
- **Homomorphic Encryption**: Homomorphic encryption enables computations to be performed directly on encrypted data, eliminating the need to decrypt sensitive information before processing. This ensures that model updates remain secure throughout training, preventing unauthorized access and inference attacks.
- **Secure Multi-Party Computation (SMPC)**: SMPC allows multiple parties to collaboratively compute functions over their inputs while keeping their individual data private. In FL, this ensures that model updates are aggregated securely without revealing individual device contributions, protecting against both inference and poisoning attacks.
- **Blockchain for FL**: Blockchain technology can enhance the transparency and integrity of FL by maintaining a decentralized and immutable ledger of model updates. By leveraging smart contracts and cryptographic verification, blockchain ensures that only authorized updates are included in the model, preventing unauthorized modifications and tampering.

Applications: FL combined with edge computing has widespread applications across various industries, improving efficiency, security, and decision-making in real-world scenarios.

Future Prospects: Federated Learning in edge computing is a rapidly evolving field, with ongoing research focused on improving scalability, energy efficiency, and security. As edge devices become more powerful and network connectivity improves, FL will play an increasingly vital role in AI-driven applications. Future advancements aim to optimize resource allocation, reduce computational overhead, and enhance secure aggregation techniques to mitigate adversarial threats. Additionally, integrating FL with emerging technologies such as 6G networks, quantum computing, and advanced cryptographic methods will further strengthen its capabilities.

IV. Adapting Explainability Techniques for Federated Learning Environments

Federated Learning (FL) is an emerging decentralized machine learning paradigm where data remains local on edge devices, and only model updates are shared. While this improves privacy, it complicates the interpretability of models due to the heterogeneity of data, models, and edge devices.

Explainable AI (XAI) plays a crucial role in building trust, accountability, and compliance (e.g., GDPR), especially in sensitive applications like healthcare, finance, and autonomous systems. However, most existing explainability techniques are designed for centralized ML, not the distributed, privacy-preserving structure of FL.

Traditional explainability tools rely on access to the complete dataset and model to generate global or local explanations. However, in FL, the model is fragmented across clients with non-IID (non-independent and identically distributed) data, personalized models, and varying computational capacities. This heterogeneity makes it difficult to generate a unified or consistent interpretation of model behavior. Additionally, directly sharing explanation artifacts, such as feature attributions or visual saliency maps, can inadvertently compromise privacy by leaking sensitive information about local datasets. Therefore, explanation techniques must be both privacy-preserving and lightweight to suit edge devices with limited bandwidth and processing power.

Several promising adaptations have been proposed to address these challenges. For instance, local explanations can be computed on edge devices using LIME or SHAP and then aggregated in a privacy-aware manner using differential privacy or secure multiparty computation. Surrogate models, such as interpretable decision trees, can also be trained locally to provide insights into complex model decisions without revealing the underlying architecture. Moreover, explainability can be integrated directly into the federated optimization process, using metrics such as feature importance divergence to align global and local interpretability objectives.

Table 1: Challenges in Applying Explainability to Federated Learning

Challenge	Explanation
Data Heterogeneity	Edge devices have non-IID (non-identically and independently distributed) data, making global explanations inconsistent across participants.
Model Diversity	Personalized federated models vary across devices; explainability needs to adapt to these local variations.
Communication Overhead	Transferring explanation artifacts (e.g., SHAP values, saliency maps) increases bandwidth and computational costs.
Security and Privacy	Explainability techniques can inadvertently leak sensitive local data through surrogate models or gradients.
Scalability	Applying traditional XAI tools on a large network of edge devices poses computational and integration challenges.

Another effective approach involves the use of federated explainability frameworks, which allow each client to compute explanations locally and contribute only the essential, obfuscated information to a centralized explanation model. Such frameworks ensure that the transparency of AI systems is maintained without sacrificing the confidentiality of user data. These methods are particularly relevant in domains such as healthcare, finance, and smart cities, where both data privacy and interpretability are critical.

In conclusion, adapting explainability techniques for federated learning environments is essential for building trustworthy, secure, and responsible AI systems. By innovating privacy-preserving and edge-efficient

explainability methods, researchers can ensure that federated models remain interpretable, even in highly distributed and heterogeneous ecosystems. Continued research in this area will play a vital role in shaping the next generation of explainable, privacy-centric AI solutions in edge computing contexts.

Table 2: Current Techniques and Their Adaptation to Federated Settings

Explainability Method	Adaptation to Federated Learning	Security Considerations
SHAP (SHapley Additive Explanations)	SHAP can be adapted using federated averaging of local explanations; however, exact value computation remains costly.	Needs differential privacy to prevent attribution inference.
LIME (Local Interpretable Model-agnostic Explanations)	LIME can be run locally on edge devices; explanations can be shared in encrypted form for global interpretability.	Vulnerable to model inversion unless protected by homomorphic encryption.
Saliency Maps (e.g., Grad-CAM)	Applied locally on vision models in edge settings; useful in medical imaging or smart surveillance.	Requires secure transfer if visualizations are shared with the server.
Counterfactual Explanations	Can be generated per device to understand local decisions.	Require caution, as they may reveal edge-device decision boundaries.

Approaches for Secure Explainability in Federated Edge Environments

- **Federated Explainability Aggregation**: Aggregate explanation vectors (e.g., gradients or SHAP scores) across clients rather than raw data or models.
- **Differentially Private Explanations**: Add noise to explanations (e.g., SHAP values) before sharing to prevent data leakage.
- **Encrypted Model Interpretation**: Use secure multi-party computation (SMPC) or homomorphic encryption to compute global explanations securely.
- **Explainability-Aware Model Updates**: Incorporate explainability metrics (e.g., feature importance divergence) during federated optimization to enhance transparency.
- **Edge-Aware Surrogate Models**: Build interpretable surrogate models (e.g., decision trees) locally for explanation without exposing main model parameters.

V. PRIVACY-PRESERVING AND COMMUNICATION-EFFICIENT XAI METHODS

The rise of federated learning and edge computing has created a critical need for explainable AI (XAI) methods that do not compromise privacy or impose heavy communication costs. Traditional XAI techniques often assume centralized data access and unrestricted computation, which contradicts the fundamental goals of privacy and efficiency in distributed environments. Therefore, it becomes essential to adapt or redesign explainability techniques to meet these constraints.

- **Local Explainability with Privacy Protection**

Methods such as LIME and SHAP can be applied locally on edge devices to generate instance-specific explanations without sharing raw data. When these explanations need to be shared or aggregated, differential privacy techniques can be used to ensure that the shared information does not allow reconstruction of sensitive user data. This local-first approach ensures compliance with privacy regulations while maintaining the interpretability of the model.

- **Surrogate Models for Lightweight Interpretation**

Surrogate models—such as shallow decision trees or simple rule-based classifiers—are used to approximate the behavior of complex models. These models are inherently interpretable and, due to their low complexity, can be trained and shared efficiently across a federated system. Sharing abstracted rule sets rather than complete models drastically reduces communication overhead and keeps sensitive internal structures private.

- **Gradient-Based Visual Explanations on Edge Devices**

Techniques like Grad-CAM and integrated gradients are adapted to function locally on vision and NLP models. These methods highlight the input features most responsible for a decision, generating heatmaps or saliency maps. When necessary, these visual outputs can be compressed or encrypted before transmission, allowing communication-efficient sharing of interpretability information without exposing raw data or model weights.

- **Aggregated Explanation Sharing**

Instead of transmitting detailed explanations, edge devices can compute summary statistics of feature importance or other explanation metrics. These statistics are aggregated in a federated manner, allowing the server to form a high-level understanding of model behavior across clients. This significantly reduces bandwidth usage and prevents exposure of client-specific patterns or outliers.

- **Knowledge Distillation for Efficient Explanation Transfer**

Federated distillation allows clients to transfer knowledge in a compressed form by distilling local models into smaller, interpretable models or explanation sets. These are then averaged or integrated centrally to generate global insights. This approach balances the need for accuracy and interpretability while maintaining communication and storage efficiency.

- **Cryptographic Techniques for Secure Explanation Computation**

Privacy-preserving computation techniques like Secure Multiparty Computation (SMPC) and Homomorphic Encryption are being applied to jointly compute global explanations without revealing local data. While computationally intensive, they provide strong privacy guarantees and are promising for high-stakes applications where both explainability and confidentiality are paramount.

XAI methods that are both privacy-preserving and communication-efficient are essential in decentralized learning environments. By leveraging localized computation, surrogate modeling, differential privacy, explanation aggregation, and cryptographic techniques, researchers and practitioners can ensure transparent, secure, and scalable AI systems. Future work should continue to focus on lightweight, federated-compatible XAI tools tailored for real-time, privacy-sensitive applications.

VI. CONCLUSION

Secure Federated Learning in edge computing is crucial for privacy-preserving AI applications. While FL mitigates data exposure risks, it remains vulnerable to adversarial and privacy attacks. Implementing robust security mechanisms such as differential privacy, homomorphic encryption, blockchain, and secure aggregation can enhance FL's resilience. However, challenges such as scalability, energy efficiency, and adversarial robustness need further investigation. This paper provides a comprehensive overview of current security threats and solutions, serving as a foundation for future advancements in secure FL for edge computing.

REFERENCES

[1] Duan, Qiang & Hu, Shijing & Deng, Ruijun & Lu, Zhihui & Yu, Shui. (2023). Combining Federated Learning and Edge Computing Toward Ubiquitous Intelligence in 6G Network: Challenges, Recent Advances, and Future Directions. IEEE Communications Surveys & Tutorials. 25(4). 1-2.

[2] Mishra, Sambit & Kumar, Nehal & Rao, Bhaskar & Brahmendra, Brahmendra & Teja, Lakshmana. (2023). Role of federated learning in edge computing: A survey. Journal of Autonomous Intelligence. 7(1).

[3] Brecko, Alexander & Kajáti, Erik & Koziorek, Jiri & Zolotová, Iveta. (2022). Federated Learning for Edge Computing: A Survey. Applied Sciences. 12(18). 1-4.

[4] Zhang, Zhuangzhuang & Wu, Libing & Ma, Chuanguo & Li, Jianxin & Wang, Jing & Wang, Qian & Yu, Shui. (2022). LSFL: A Lightweight and Secure Federated Learning Scheme for Edge Computing. IEEE Transactions on Information Forensics and Security. PP(99). 1-2.

[5] Wang, J., Liu, X., Liang, W., et al. (2022). "Blockchain-Based Federated Learning for Secure Edge Computing." *IEEE Transactions on Services Computing*, 15(1), 176-189.

[6] Kairouz, P., McMahan, H. B., Avent, B., et al. (2021). "Advances and Open Problems in Federated Learning." *Foundations and Trends in Machine Learning*, 14(1-2), 1-210.

[7] Emmanni, Phani Sekhar. (2021). Federated Learning for Cybersecurity in Edge and Cloud Computing Federated Learning for Cybersecurity in Edge and Cloud Computing. International Journal of Computing and Engineering. 2(1). 14-25.

[8] Liu, Gaoyang & Wang, Chen & Ma, Xiaoqiang & Yang, Yang. (2021). Keep Your Data Locally: Federated Learning-Based Data Privacy Preservation in Edge Computing. IEEE Network. 35(2). 60-66.

[9] Xu, J., Zhou, F., & Wang, R. (2021). "Security and Privacy in Federated Learning: A Survey." *Future Generation Computer Systems*, 115, 466-480.

[10] Zhao, Z., Sun, H., Zhang, T., et al. (2021). "Privacy-Preserving Federated Learning for IoT Edge Computing." *IEEE Internet of Things Journal*, 8(2), 1157-1171.

[11] Liu, Y., Yang, Q., Jiang, X., & Xiong, L. (2021). "Secure Aggregation for Federated Learning in Edge Computing." *IEEE Transactions on Big Data*, 7(1), 110-124.

[12] Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). "Federated Learning: Challenges, Methods, and Future Directions." *IEEE Signal Processing Magazine*, 37(3), 50-60.

[13] Ye, Yunfan & Li, Shen & Liu, Fang & Tang, Yonghao & Hu, Wanting. (2020). EdgeFed: Optimized Federated Learning Based on Edge Computing. IEEE Access. 8(1). 209191-209198.

[14] Ma, X., Zhou, Z., Chen, J., et al. (2020). "Safeguarding Federated Learning in Edge Computing with Differential Privacy." *IEEE Transactions on Neural Networks and Learning Systems*, 31(12), 5373-5385.

[15] Bonawitz, K., Eichner, H., Grieskamp, W., et al. (2019). "Towards Federated Learning at Scale: System Design." *Proceedings of the 2nd SysML Conference*.

[16] Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). "Federated Machine Learning: Concept and Applications." *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2), 1-19.

[17] Hard, A., Rao, K., Mathews, R., et al. (2019). "Federated Learning for Mobile Keyboard Prediction." *arXiv preprint arXiv:1811.03604*.