People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research
KASDI MERBAH UNIVERSITY - OUARGLA
Faculty of New Technologies of Information and Telecommunication
Department of Computer Science and Information Technology

MASTER THESIS

Domain: Computer Science
Field: Artificial Intelligence and Data Science

Thesis Title

Towards Privacy-Preserving Federated Learning with Explainable AI:
Applications in IoT Networks

Submitted by:
Khouildat Houria
Hideb Hadjer

Supervised by:
Mr. Belal Khaldi

Date: 10/06/2025

Before the Jury:

Name	Role	Affiliation
Khaldi Belal	Supervisor	UKM Ouargla
Bensaci Remla	President	UKM Ouargla
Aiadi Oussama	Examiner	UKM Ouargla

Academic Year: 2024/2025

Abstract

This research aims to explore and develop an integrated framework that combines **Federated Learning (FL)** with **Explainable Artificial Intelligence (XAI)**, with the objective of enhancing data privacy and achieving transparency in AI decision-making processes, particularly within sensitive environments such as **Internet of Things (IoT)** networks and **Intelligent Medical Systems (IoMT)**. The study is motivated by the growing need for intelligent solutions that not only ensure high performance but also respect ethical and legal standards regarding data confidentiality and users' right to understand automated decisions.

The work addresses the core challenges faced by federated learning, especially the difficulty of applying traditional interpretability techniques due to the decentralized nature of data and the lack of unified access to a global model. It also discusses the inherent tension between applying strong privacy-preserving techniques (such as differential privacy and encryption) and the need for human-understandable interpretations that clarify the model's reasoning process.

In this context, the study adopts a convolutional neural network (CNN) architecture deployed in a federated learning environment, evaluating a range of prominent FL strategies (e.g., FedAvg, FedOpt, FedAdam) across different scenarios involving varied data distributions (balanced, random, skewed), synchronization modes (synchronous and asynchronous), and different numbers of participating clients.

Special attention is given to integrating visual interpretability tools like **GradCAM** and **NormGrad**, which provide insight into the key input regions influencing the model's predictions. The proposed framework also considers performance-related aspects such as computational efficiency, communication cost, number of message exchanges, training time, and privacy protection.

This project aspires to contribute toward the design of decentralized, interpretable AI models capable of balancing predictive accuracy with data protection and ethical transparency—laying the foundation for responsible applications in fields such as medicine, cybersecurity, and financial services.

Keywords

Federated Learning, Privacy Preservation, Explainable Artificial Intelligence (XAI), IoT Networks, Performance Evaluation in Federated Learning, Model Interpretability, Data Security, Communication Efficiency in Federated Learning, Message Exchange, Trade-off Between Performance and Privacy.

الملخص

يهدف هذا البحث إلى استكشاف وتطوير إطار متكامل يجمع بين **التعلم الفيدرالي والذكاء الاصطناعي القابل للتفسير**، وذلك من أجل تعزيز حماية الخصوصية وتحقيق الشفافية في قرارات أنظمة الذكاء الاصطناعي، لا سيما ضمن البيئات الحساسة مثل **شبكات إنترنت الأشياء (IOT) والأنظمة الطبية الذكية (IOMT)**.

تأتي هذه الدراسة في ظل الحاجة المتزايدة إلى حلول ذكية لا تكتفي بتحقيق الأداء العالي، بل تراعي كذلك الجوانب الأخلاقية والتشريعية المتعلقة بسرية البيانات وحق المستخدم في فهم قرارات الأنظمة التنبؤية.

يناقش البحث التحديات الجوهرية التي تواجه التعلم الفيدرالي، لا سيما صعوبة تطبيق تقنيات التفسير التقليدية نظرًا للطبيعة اللامركزية للبيانات وغياب الوصول الكامل إلى النموذج العالمي أو البيانات الموحدة.

كما يتناول التوتر القائم بين الحاجة إلى حماية البيانات عبر تقنيات مثل الخصوصية التفاضلية والتشفير، مقابل الحاجة إلى تفسير قابل للفهم يتيح تتبع منطق اتخاذ القرار داخل النموذج.

في هذا السياق، يعتمد البحث على بنية أساسية تتكون من شبكة عصبية التفافية (CNN) ضمن إطار تعلم فيدرالي، مع اختبار مجموعة من الاستراتيجيات الفيدرالية الشهيرة (مثل FedAvg، Fedopt، FedAdam وغيرها) تحت سيناريوهات متنوعة تشمل اختلاف توزيع البيانات (متساوي، عشوائي، غير متوازن)، واختلاف آليات التزامن (متزامن وغير متزامن)، وعدد العملاء المشاركين في التدريب.

كما يولي البحث اهتمامًا خاصًا بتكامل تقنيات التفسير البصري مثل **GradCAM** و**NormGrad**، التي تمثل أدوات فعالة لتسليط الضوء على المناطق الأكثر تأثيرًا في المدخلات التي يعتمد عليها النموذج عند اتخاذ القرار.

ويُناقش الإطار المقترح أيضًا أبعادًا متعددة للأداء تشمل: الكفاءة الحسابية، استهلاك الاتصال، عدد الرسائل المتبادلة، زمن التدريب، وحماية الخصوصية.

يسعى هذا المشروع إلى المساهمة في بناء نماذج ذكاء اصطناعي لامركزية وأكثر شفافية، قادرة على التوفيق بين الكفاءة التنبؤية والامتثال لمبادئ حماية الخصوصية والعدالة التفسيرية، مما يمهد الطريق نحو تطبيقات أكثر مسؤولية وثقة في قطاعات حيوية مثل الطب، الأمن السيبراني، والخدمات المالية.

الكلمات المفتاحية

التعلم الفيدرالي، الحفاظ على الخصوصية، الذكاء الاصطناعي القابل للتفسير (XAI)، شبكات إنترنت الأشياء، تقييم الأداء في التعلم الفيدرالي، تفسير نماذج الذكاء الاصطناعي، أمن البيانات، كفاءة الاتصال في التعلم الفيدرالي، تبادل الرسائل، الموازنة بين الأداء والخصوصية.

Contents

Abstract **2**

الملخص **4**

1 General Introduction **11**
1.1 Background and Motivation 11
1.2 Problem Statement 20
1.3 Research Objectives 23
1.4 Contributions of the Study 23
1.5 Thesis Structure 24

2 Related Work **26**
2.1 Privacy and Security in Federated Learning 26
2.2 Explainable AI and Federated Learning 31
2.3 Federated Learning in IoT and Healthcare 33
2.4 Federated Learning in Advanced Systems 35

3 Proposed method **39**
3.1 The Overall Framework of the Proposed Approach 39
3.1.1 Data Collection Stage 40
3.1.2 Local Model Training Stage 41
3.1.3 Interpretability Mechanisms 42
3.1.4 Central Aggregation Phase (Aggregation Server) 44
3.1.5 Prediction and Interpretation 48

4 Experimental evaluation and discussion **50**
4.1 Data Preparation 50
4.2 Evaluation Metrics 51
4.3 A Comparative Study of Federated Learning Strategies (FedAvg, FedAdam, FedYogi, FedOpt) 52
4.4 Analysis of Data Distribution Impact on FedAvg Performance 56
4.5 The effect of Sync vs. Async on the FedAvg algorithm 59
4.6 Interpreting Models Using GradCAM and NormGrad 63

4.7 The Impact of Privacy-Preserving Noise on Model Performance and Interpretability in Federated Learning 65
4.8 Impact of Model Interpretability using GradCAM and NormGrad in Standard and Multi-task Federated Learning 69
4.9 Investigating Weight Sharing Effectiveness in Federated Learning Under Different Data Distributions 71
4.10 Impact of Model Interpretability using GradCAM and NormGrad 73
4.11 Impact of the Number of Clients on Performance and Interpretability . . . 76
4.12 Physical Resources and Federated Learning Metrics 79

List of Figures

1.1 Federated Learning Process adapted from [3]. 12
1.2 representative figure on how may explainability promote trust between man and machine 17
1.3 Explainability techniques 18
1.4 Relationship between Federated learning and Interpretability 19

3.1 Interpretable Federated Learning Architecture for Healthcare 40
3.2 GradCAM method 43
3.3 NormGrad and GradCAM Process 44

4.1 Photo of data 50
4.2 accuracy plot of FedAvg 52
4.3 local accuracy of client in FedAvg 53
4.4 plot local accuracy of client in FedAvg 53
4.5 accuracy plot of FedAdam 53
4.6 local accuracy of client in FedAdam 53
4.7 plot local accuracy of client in FedAdam 53
4.8 accuracy plot of FedOpt 54
4.9 local accuracy of client in FedOpt 54
4.10 plot local accuracy of client in FedOpt 54
4.11 accuracy plot of FedYogi 54
4.12 local accuracy of client in FedYogi 55
4.13 plot local accuracy of client in FedYogi 55
4.14 Accuracy of Equal,Random and Unequal Sync in the FedAvg 57
4.15 client accuracy in equal sync 57
4.16 plot client accuracy in equal sync 57
4.17 client accuracy in random sync 57
4.18 plot client accuracy in random sync 57
4.19 client accuracy in unequal sync 58
4.20 plot client accuracy in unequal sync 58
4.21 Accuracy of Sync and Async in the FedAvg 60
4.22 client accuracy in sync 60
4.23 client accuracy in async 61

4.24 client accuracy in async . . . 61
4.25 Interpretation of image 1 . . . 63
4.26 Interpretation of image 2 . . . 63
4.27 Interpretation of image 3 . . . 63
4.28 Interpretation of image 4 . . . 64
4.29 Interpretation of image 5 . . . 64
4.30 Accuracy during Noise Level . . . 66
4.31 Privacy Score during Noise Level . . . 66
4.32 Interpretability in 0 Noise Level . . . 67
4.33 Interpretability in 0.1 Noise Level . . . 67
4.34 Interpretability in 0.5 Noise Level . . . 67
4.35 Interpretability in 1 Noise Level . . . 67
4.36 accuracy of Standard and Multi-task Federated Learning . . . 69
4.37 Interpretation of Standard Federated Learning . . . 70
4.38 Interpretation of task Federated Learning . . . 70
4.39 accuracy plot . . . 72
4.40 accuracy plot Net(1) . . . 73
4.41 accuracy plot Net(2) . . . 74
4.42 Interpretability Net(1) . . . 74
4.43 Interpretability Net(2) . . . 74
4.44 plot of accuracy using different numbers of clients: 2, 5, 10, and 20 . . . 76
4.45 Interpretation of a sample image for Client 2. . . . 77
4.46 Interpretation of a sample image for Client 5. . . . 77
4.47 Interpretation of a sample image for Client 10. . . . 77
4.48 Interpretation of a sample image for Client 20. . . . 77

List of Tables

1.1 A comparison between different architectures of FL: Centralized, Semi-Centralized, and Fully Decentralized. 15

2.1 Summary of Selected Studies 1–20 37

4.1 Detailed performance summary of federated strategies with training and evaluation accuracy 55
4.2 Comparison of FedAvg Performance Across Different Data Distribution Types 58
4.3 Summary of Client Training Accuracy Experiments for Sync and Async Strategies 62
4.4 Visual Interpretation of the CNN Model's Attention 64
4.5 Impact of Privacy Noise on Model Accuracy and Visual Explanations . . . 68
4.6 Qualitative comparison between standard and multi-task federated learning using GradCAM and NormGrad. 70
4.7 Accuracy and Weight Sharing Observations Across Data Distributions . . . 72
4.8 Comparison of simple and complex models using GradCAM and NormGrad. 75
4.9 Impact of client number on model performance and heatmap interpretability. 78
4.10 Final Summary of Federated Learning Performance Metrics 79

CHAPTER 1

Chapter 1

General Introduction

1.1 Background and Motivation

Artificial Intelligence (AI) has witnessed tremendous development in recent years, leading to its widespread use across various sectors such as healthcare, finance, cybersecurity, and the Internet of Things (IoT). However, most traditional AI systems rely on aggregating data in a centralized location for model training, which raises several concerns related to privacy, security, and regulatory compliance. To address these issues, Federated Learning has emerged as a new approach aimed at enhancing data privacy during the training of intelligent models.

Federated Learning (FL) is a decentralized machine learning technique that enables multiple devices or servers to train a shared model without the need to exchange raw data between them [1]. Instead, the model is trained locally on each device, and only the parameter updates (such as weights and gradients) are sent to a central server, which aggregates them to create an improved global model.

Unlike traditional centralized learning, where data is collected in a central server, federated learning allows edge devices such as smartphones, Internet of Things (IoT) devices, hospitals, and banks to contribute to model training while preserving their data privacy [1]. This approach is particularly important in privacy-focused machine learning, as it reduces the risk of sensitive data leakage and ensures compliance with regulations such as the General Data Protection Regulation (GDPR) in the European Union and the Health Insurance Portability and Accountability Act (HIPAA) in the United States [2]. Moreover, training models without the need to transfer data minimizes transmission costs and enhances performance efficiency, especially in environments with limited or slow internet connectivity.

The federated learning mechanism follows these key steps:

- **Initial model distribution:** A pre-trained AI model is sent to participating devices.
- **Local training:** Devices train the model using their local data without sending

the data to any external entity.

- **Sending updates:** Only the optimized parameters (such as weights and gradients) are sent to the central server.
- **Aggregating updates:** The central server aggregates updates from various devices to refine the global model.
- **Repeating the process:** The updated model is redistributed to devices for further training, and the process continues until an optimal and accurate model is achieved.

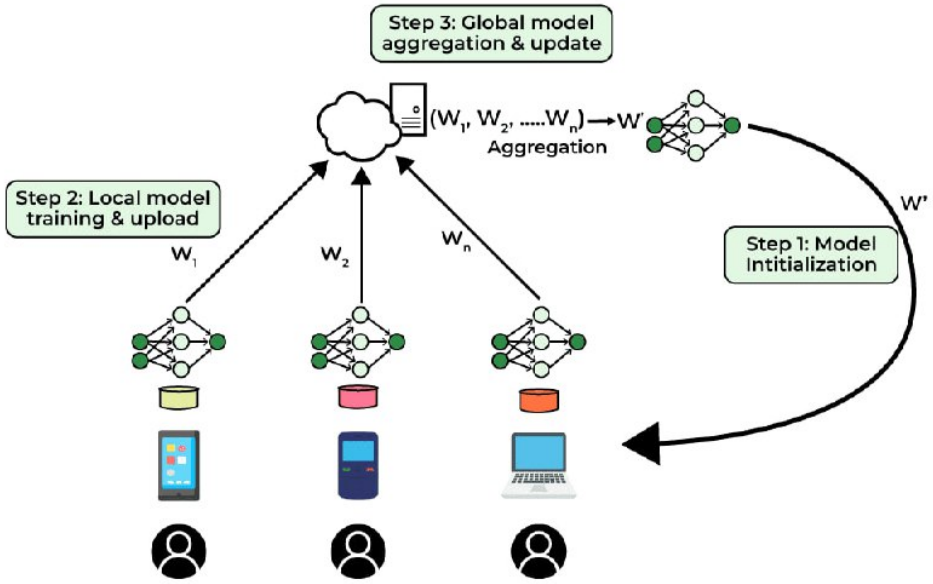

Figure 1.1: Federated Learning Process adapted from [3].

Federated learning has vast potential across various fields that require a balance between artificial intelligence and data privacy. Some of the most notable applications include:

Healthcare

Federated learning enables hospitals and healthcare institutions to train AI models based on patient data without transferring that data to external centers. This helps improve medical diagnostics, develop personalized treatments, and analyze the progression of chronic diseases while ensuring patient privacy [2]. For example, it can be used to analyze X-ray images for disease diagnosis without sharing patient data between hospitals, thereby achieving a balance between model accuracy and data protection .

Finance

Banks and financial institutions use federated learning to detect financial fraud and analyze customer behavior without centralizing their data. This approach reduces security

risks and ensures transaction confidentiality [3]. For instance, a financial institution can leverage federated learning to identify suspicious transactions across multiple clients without compromising their privacy, thus enhancing the overall security of payment systems .

Edge Computing

Federated learning strengthens edge computing by enabling smart devices, such as mobile phones, to learn from user data locally without sending it to cloud servers. This technique is used to personalize user experiences and improve smart application performance [4]. For example, Google's Gboard keyboard utilizes federated learning to enhance word prediction accuracy based on user input without transmitting data to the internet, thereby improving both user privacy and processing efficiency .

Cybersecurity

Federated learning enhances cybersecurity by improving malware detection and cyberattack defense systems. Instead of sharing user data, different systems can train models locally to identify cyber threats and analyze attack patterns while maintaining privacy [5]. For example, cybersecurity companies can use federated learning to analyze malware spread across multiple devices, helping to develop stronger defenses without compromising user privacy or sharing sensitive data .

Federated Learning can be classified into three main types based on the architecture used: Centralized FL, Semi-Centralized FL, and Fully Decentralized FL. Each of these architectures differs in terms of structure, performance, privacy, and security challenges.

In Centralized Federated Learning (Centralized FL), a single central server is responsible for coordinating the training process. Clients (edge devices) train models locally using their private data and then send updates to the central server, which aggregates them to improve the global model. This approach is simple to implement and allows precise monitoring of the central model [17]. However, relying on a central server creates a single point of failure, making it vulnerable to cyberattacks and connectivity issues. Additionally, it may consume high bandwidth when handling a large number of connected devices [19].

Semi-Centralized Federated Learning (Semi-Centralized FL) serves as a middle ground between centralized and fully decentralized systems. In this model, multiple edge servers act as intermediaries between clients and the central server. These edge servers aggregate updates from a group of local devices before sending them to the central server. This architecture reduces the load on the central server, minimizes bandwidth consumption, and improves fault tolerance compared to the fully centralized approach [23]. However, security risks still exist since edge servers can become targets for cyberattacks, and task distribution increases the complexity of coordination among devices [3].

On the other hand, Fully Decentralized Federated Learning (Fully Decentralized FL) is

the most secure and privacy-preserving approach, as there is no central server. Instead, updates are exchanged directly between devices in a peer-to-peer (P2P) manner using protocols like Blockchain or Gossip Learning. This method allows model updates to be shared among clients without requiring an external server [5]. This architecture ensures maximum security and privacy by preventing data transmission to an external entity and eliminating a single point of failure, making it ideal for sensitive applications such as cybersecurity and Internet of Things (IoT) systems. However, this approach requires advanced communication protocols and more complex coordination mechanisms, making it more challenging to implement and manage compared to the other two architectures.

Table 1.1: A comparison between different architectures of FL: Centralized, Semi-Centralized, and Fully Decentralized.

Feature	**Centralized FL**	**Semi-Centralized FL**	**Fully Decentralized FL**
Architecture	A central server coordinates model training.	Multiple edge servers aggregate updates before sending to a central server.	No central server; nodes communicate directly with each other.
Communication Flow	Clients send model updates to a central server, which aggregates and updates the global model.	Clients send updates to edge nodes, which partially aggregate updates before sending them to a central server.	Peers communicate directly, exchanging updates in a peer-to-peer (P2P) fashion.
Scalability	Limited by the server's capacity and bandwidth.	More scalable than centralized due to distributed aggregation.	Highly scalable, as there is no central bottleneck.
Latency	Higher due to dependence on a single server.	Moderate, as edge nodes reduce direct server load.	Lower, as updates are exchanged locally.
Privacy & Security	Model updates are sent to a central server, creating a single point of failure.	Edge nodes add an extra layer of security but still rely on a central server.	Higher privacy, as no central entity collects updates.
Fault Tolerance	Low—failure of the central server disrupts the entire system.	Medium—failure of an edge node impacts some clients but not the entire system.	High—no single point of failure, as peers can reconfigure dynamically.
Computational Overhead	The central server requires high computational resources for model aggregation.	Computational load is shared between the central server and edge nodes.	Computationally efficient as nodes contribute to training and aggregation.
Deployment Complexity	Easier to implement and manage.	More complex due to additional edge nodes.	Most complex due to lack of central coordination and reliance on P2P protocols.
Example Applications	Healthcare institutions collaborating with a central hospital.	Smart cities with multiple edge computing layers.	Blockchain-based FL, distributed IoT networks.

Understanding these architectures is crucial, especially when comparing them to traditional centralized learning approaches, which differ significantly in terms of data privacy and model training efficiency such that:

Data Privacy

In centralized learning, all data is transferred to a central server, where it is processed and used to train models [1]. However, this aggregation poses significant risks related to privacy breaches, as the data becomes more susceptible to security attacks, privacy leaks, and violations of regulatory frameworks such as (GDPR) and (HIPAA) [15]. Studies like that of Shokri & Shmatikov (2015) have shown that data sharing in centralized learning can lead to re-identification of users' data through privacy attacks [23].
In contrast, federated learning keeps data on local devices, and instead of transferring raw data, only model updates are shared, which reduces the risk of data leakage and enhances compliance with data protection regulations [3]. For instance, in medical applications, multiple hospitals can train an AI model without sharing sensitive patient data, thereby preserving privacy and minimizing security risks [2]. Additionally, studies such as Abadi et al. (2016) have demonstrated that techniques like Differential Privacy can be integrated with federated learning to further enhance security [22].

Model Training

In centralized learning, all data is gathered into a single server, allowing the model to leverage a complete dataset, which often results in higher accuracy, especially when large amounts of data are available [6]. However, this approach requires substantial computational power and high-speed data transfer, which can lead to massive resource consumption and latency issues when dealing with extensive datasets [17]. Moreover, if the central server fails, the entire training process halts, making it less fault-tolerant [7].
On the other hand, in federated learning, models are trained in a distributed manner across multiple devices, reducing the need for data transmission and central resource consumption [1]. However, this approach faces challenges such as heterogeneity in data across devices, which may lead to reduced accuracy if not handled properly [8]. Additionally, poor network connections or device performance variations can impact training efficiency [13]. To address these challenges, aggregation techniques such as Federated Averaging (FedAvg) are employed to efficiently update the global model [17].

On the other hand, with the increasing reliance on artificial intelligence in sensitive fields, decision explainability has become a fundamental issue to ensure trust in these systems. Explainability refers to a model's ability to clarify how and why a specific decision was made, allowing users to validate predictions and make informed decisions accordingly [6].

explainability techniques play a crucial role in enhancing transparency and fostering

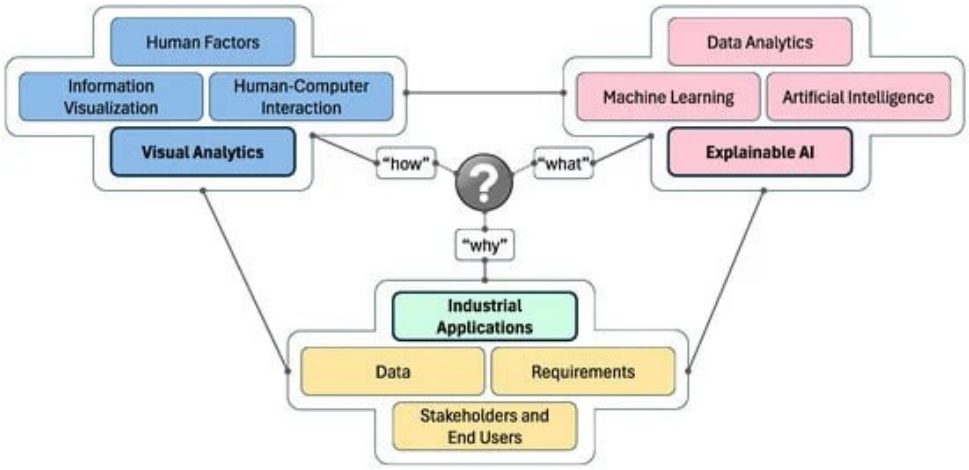

Figure 1.2: representative figure on how may explainability promote trust between man and machine

trust in model decisions. These techniques can be categorized into several main types based on the interpretation method used. First, feature importance-based approaches, such as SHAP, LIME, and Integrated Gradients, analyze the influence of each feature on the model's output, helping to identify the most impactful factors in decision-making [6]. Second, visualization-based techniques, including Grad-CAM, Saliency Maps, and Occlusion Sensitivity, provide visual representations of how the model focuses on specific areas of the data during predictions, making them particularly valuable for computer vision applications [25].

Additionally, surrogate models, such as decision trees and linear models, are used to approximate the behavior of complex models, offering more transparent explanations of how decisions are made [26]. Decision-based methods, such as Counterfactual Explanations and Example-Based Explanations, provide deeper insights into how small input modifications affect model predictions, allowing users to understand decision logic more precisely [27].

However, applying these techniques in a federated learning environment presents challenges related to data distribution and the lack of direct access to complete information.

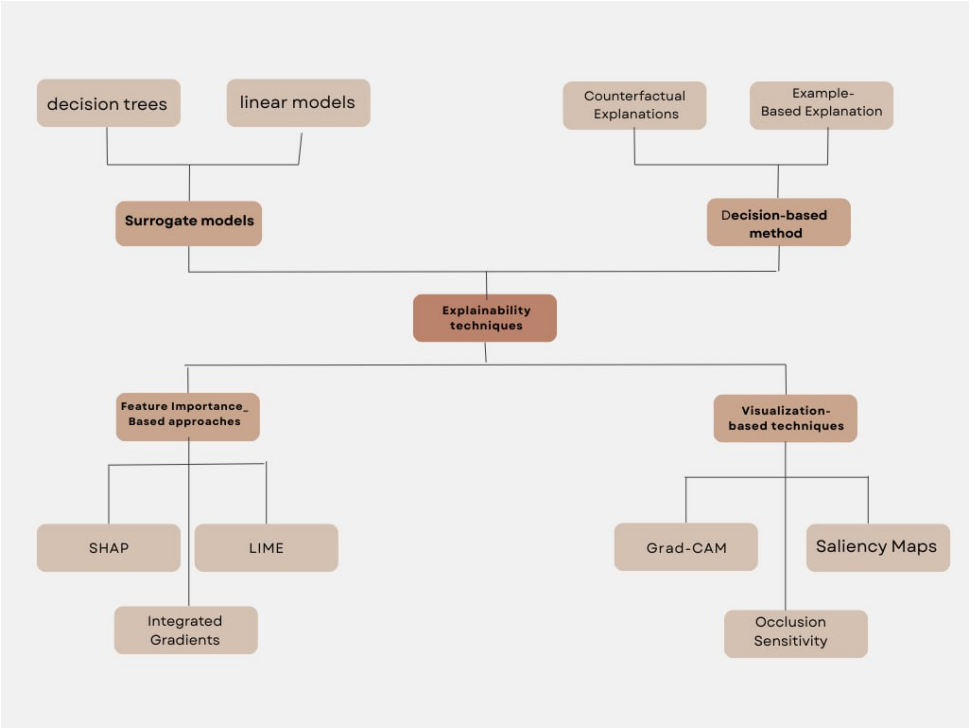

Figure 1.3: Explainability techniques

Despite the importance of explainability, its implementation in systems faces several key challenges:

Model Complexity

Traditional machine learning relies on complex models such as deep neural networks, which are often considered ”black boxes,” making it difficult to understand how decisions are made [6]. For example, in disease diagnosis using machine learning, it may be difficult to explain why a particular X-ray image was classified as having a tumor, due to the lack of clear interpretation of the factors that led to this classification .

Inconsistency in Interpretations

The explainability results may vary across different devices, making it difficult to standardize explanations to ensure fairness and transparency.

Example: An AI system running on smartphones may provide different explanations for the same decision depending on the data available on each device, leading to inconsistent results and reduced user trust in the model.

As we previously learned, federated learning aims to strike a balance between intelligent data analysis and protecting user privacy, while interpretability techniques seek to make model decisions more transparent and explainable. However, achieving both goals simultaneously presents a challenge, as some federated learning principles conflict with the need to access model data for interpretability purposes.

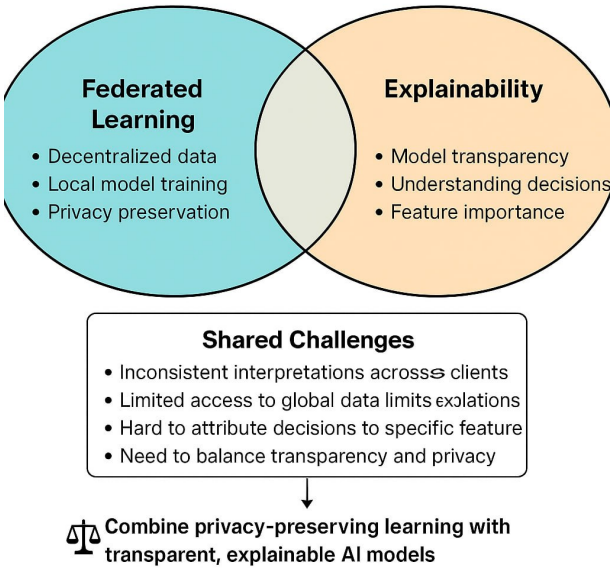

Figure 1.4: Relationship between Federated learning and Interpretability

Because federated learning protects data privacy and ensures interpretability, unlike traditional centralized learning, where all data is easily accessible and analyzed, federated learning keeps data distributed across multiple machines, making it more difficult to interpret model decisions. The main challenges can be summarized as follows:

Lack of Centralized Data for Interpretation

In traditional learning, researchers can analyze all data together to understand how each feature influences the model's decisions. However, in federated learning, data remains distributed across different devices, preventing centralized analysis [3].

Example : In a healthcare system using federated learning to analyze X-ray images, it is not possible to combine data from all hospitals in one place to examine common patterns, making it difficult to interpret model decisions globally .

Balancing Privacy and Explainability

Some explainability techniques require access to raw data or internal model parameters, which contradicts the principles of federated learning that aim to protect user privacy. Techniques such as SHAP and LIME rely on reconstructing partial data to understand the impact of each variable on the outcome, which may lead to risks related to re-identifying users [8].

Example: In a medical environment, explaining decisions made by a federated model using patient data may require comparing different medical records, which is not possible without violating privacy principles [9].

Interpreting model decisions plays a fundamental role in ensuring the reliability of AI systems, particularly in federated learning environments . Understanding how these models make decisions helps achieve the following objectives:

Enhancing Trust in Intelligent Systems

The more users and system developers understand how decisions are made, the more trust they will have in these systems. In federated learning, where users cannot see all the data, decision explainability becomes essential to ensure that the model does not behave in a biased orunexpected manner[9].
Example: In medical classification systems, understanding why a specific disease is diagnosed based on an X-ray helps doctors make more reliable treatment decisions .

Improving Performance

By interpreting the errors made by a model, weaknesses can be identified and corrected, thereby improving prediction accuracy [10].
Example: If a federated model for fraud detection in banking makes inaccurate decisions due to irrelevant features, developers can use explainability techniques to identify features negatively impacting performance and retrain the model more efficiently .

Ensuring Regulatory Compliance

Many laws and regulations, such as (GDPR) and (HIPAA) , require AI system decisions to be interpretable [11].
Example: If a loan is denied based on a federated model's decision, the bank must provide a clear explanation of the reason; otherwise, it may face legal accountability .

1.2 Problem Statement

After an in-depth study, these challenges were identified as major obstacles to the broader adoption of Federated Learning, especially in scenarios where a clear understanding of model behavior and decision-making processes is essential. of Existing challenges in FL model transparency:

Lack of Centralized Access to Global Data

In a federated learning (FL) environment, data remains distributed across multiple devices, preventing centralized access to all data. This makes it difficult to apply traditional analysis and interpretability techniques, which rely on an integrated dataset [1]. Additionally, data distribution may lead to challenges in balancing model accuracy and interpretability, particularly in systems that require high transparency, such as healthcare and finance [2].

Difficulty in Applying Standard Interpretability Techniques in FL

Most model interpretability techniques depend on analyzing the entire dataset and model, which is not feasible in FL due to privacy requirements. For example, techniques such as

LIME and SHAP require analyzing the impact of each feature within a complete dataset, which is unavailable in FL due to its distributed nature [3]. Furthermore, the lack of direct access to updated client parameters complicates the interpretability process [4].

Beyond the technical challenges of model interpretability in Federated Learning, there are fundamental issues concerning trust and the reliability of FL decisions. The lack of transparency in decision-making processes, coupled with limited traceability of distributed model updates, weakens user trust and hinders the adoption of FL in critical domains that require fairness and accountability.

Trust Issues in Federated Learning

- **Data Privacy and Security**
  Although FL prevents direct data sharing, model updates may still expose sensitive information through gradient leakage or model inversion attacks. Users may hesitate to participate in FL if strong guarantees for data protection are not provided [17].

- **Security Attacks and Model Robustness**
  Malicious clients can launch Byzantine attacks by submitting compromised model updates, negatively impacting overall model performance. Additionally, model poisoning attacks allow adversaries to manipulate inputs or updates to achieve specific outcomes, undermining model reliability [18].

- **Fairness and Bias in FL Models**
  Heterogeneous data distribution across clients can introduce bias in the model's decisions, leading to unfair outcomes for certain groups. Unequal data representation affects model reliability and reduces trust in critical applications like healthcare and finance [12].

- **Accountability and Lack of Transparency**
  Since FL relies on distributed training, it becomes difficult to assign responsibility for biased decisions or errors. The absence of proper tracking and auditing mechanisms lowers regulatory confidence in deploying FL for banking, medical, and other sensitive domains [19].

Explainability Issues in Federated Learning

- **Lack of Model Decision Transparency**
  Most FL models leverage deep neural networks, which function as black boxes, making it difficult to understand how decisions are made. This limits FL adoption in fields that require decision justification, such as medical diagnosis or financial approvals [7].

- **Opaque Model Updates**
  Participants cannot determine how their local updates contribute to the global

model, as FL aggregates updates across all clients. This reduces the ability to detect biased or adversarial clients, potentially corrupting the model [20].

- **Local Data Distribution Limits Explainability**
  Unlike centralized learning, researchers in FL lack access to the entire dataset, making it harder to apply traditional interpretability techniques such as SHAP or LIME. This limits the ability to analyze feature importance in model predictions [21].

As Federated Learning continues to gain traction in sensitive environments such as healthcare and finance, the need for novel interpretability techniques that maintain data privacy and model efficiency becomes increasingly critical. Striking a balance between explainability and privacy is essential to ensure transparency and foster user trust. With the increasing adoption of Federated Learning (FL) in sensitive environments, there is a growing need for new mechanisms that enhance explainability without compromising data privacy or model performance. While FL limits direct data sharing, its lack of a holistic view of the data makes model decision interpretation more complex.

On the other hand, privacy-preserving techniques such as homomorphic encryption and differential privacy may reduce model accuracy or hinder interpretability. Therefore, achieving a balance between transparency and protection requires the development of FL-compatible explainability techniques, such as federated explainability or privacy-preserving model analysis methods.

Innovating such solutions will not only boost user trust in FL but will also accelerate its adoption in critical domains that demand transparency and accountability.

Some statistics to support the need for improved explainability in Federated Learning include:

Growth of Federated Learning Adoption

With increasing regulations restricting the sharing of sensitive data, Federated Learning (FL) has emerged as a strategic solution for preserving privacy while training intelligent models. According to a report by IBM, the number of organizations adopting FL has grown by 35% over the past three years, highlighting a strong shift towards decentralized learning [13].

Privacy Concerns as a Key Driver for FL Adoption

Data shows that companies are becoming more sensitive to privacy issues, driving them to adopt solutions like FL. A report by the European Union indicates that 78% of large enterprises prefer FL due to its ability to protect data and reduce the risk of information leaks [14].

Impact of Explainability on AI Performance in Practical Applications

Explainability (XAI) plays a crucial role in enhancing user trust in AI models, particularly in critical domains like healthcare and finance. A study found that interpretable models improved doctors' decision accuracy by 20% when used for disease diagnosis [16].

1.3 Research Objectives

the research aims to achieve the following objectives:

- To design a unified framework that effectively integrates Federated Learning (FL) with Explainable AI (XAI), enabling decentralized model training with interpretable decision-making capabilities.
- To investigate the balance between privacy and interpretability, specifically in resource-constrained and sensitive environments such as IoT networks and medical systems, where data confidentiality and decision transparency are both critical.
- To empirically evaluate various FL strategies (e.g., FedAvg, FedAdam, FedYogi, FedOpt) under different data distributions and synchronization modes, assessing their impact on accuracy, communication cost, and interpretability.
- To explore the applicability of visual interpretability tools, such as GradCAM and NormGrad, within the federated setting to assess their effectiveness in offering insight into model decisions without breaching user privacy.
- To propose guidelines and methodologies for deploying interpretable, privacy-preserving AI systems in real-world IoT applications.

1.4 Contributions of the Study

This study makes several technical and methodological contributions to the field of privacy-aware, interpretable federated learning:

- **An end-to-end FL-XAI framework:** A comprehensive and practical system that enables privacy-preserving model training while maintaining a high level of explainability through GradCAM and NormGrad across heterogeneous clients.
- **Performance evaluation of FL strategies:** A rigorous empirical comparison of multiple federated learning algorithms (FedAvg, FedAdam, FedYogi, FedOpt) focusing solely on their performance metrics. Additionally, a detailed analysis was conducted on the FedAvg algorithm under different data distributions (equal, random, skewed) and synchronization modes (synchronous vs. asynchronous), highlighting its sensitivity to distribution and communication patterns.

- **Integration of interpretability into FL workflow:** Novel integration of heatmap-based explainability tools within the federated learning lifecycle, allowing model transparency without requiring access to raw data.
- **Quantitative analysis of the privacy-interpretability trade-off:** Demonstrating how privacy-preserving techniques (e.g., noise injection) affect the accuracy and clarity of model explanations, which is rarely explored in existing literature.
- **Scalability and efficiency insights for IoT:** Practical insights into deploying interpretable FL in bandwidth-limited and latency-sensitive environments, including performance under different client numbers and communication constraints.

1.5 Thesis Structure

This thesis is structured as follows:

- **Chapter 1: General Introduction**
  Provides an overview of the research background, outlines the problem statement, defines research objectives, highlights the study's contributions, and presents the thesis organization.
- **Chapter 2: Related Work**
  Reviews previous studies on privacy and security in FL, the integration of explainable AI in FL, and the application of FL in IoT and healthcare environments.
- **Chapter 3: Proposed Method**
  Describes the system architecture, model design, interpretability mechanisms, and implementation of various FL strategies such as FedAvg, FedAdam, FedYogi, and FedOpt.
- **Chapter 4: Experimental Evaluation and Discussion**
  Presents experimental settings, data preparation, evaluation metrics, performance comparisons across FL strategies, impact of data distribution and synchronization, integration of interpretability (GradCAM and NormGrad), and privacy-performance trade-offs.
- **Chapter 5: Conclusion and Future Work**
  Summarizes the key findings of the study, discusses limitations, and outlines potential directions for future research.

CHAPTER 2

Chapter 2

Related Work

2.1 Privacy and Security in Federated Learning

The researchers in this study [28] conducted a systematic review of the security threats facing the privacy of Federated Learning (FL) in Internet of Things (IoT) environments, with a focus on potential countermeasures to enhance security. The study identified several key threats, including inference attacks, which allow attackers to infer sensitive information from model updates, poisoning attacks, which aim to corrupt the model by feeding it misleading data, and data eavesdropping during transmission.

To address these risks, differential privacy (DP) was adopted as one of the main techniques, where noise is added to the model updates to prevent the inference of sensitive information about the participants. However, the researchers pointed out that excessive noise addition could lead to a decline in model accuracy, creating a challenge in achieving an optimal balance between privacy and performance. Homomorphic encryption (HE) was also employed to allow computations to be performed directly on encrypted data without the need for decryption, providing a high level of security. However, the high computational cost of this technique is a major barrier, especially for low-power devices in IoT environments.

Additionally, secure aggregation (SA) was used to prevent information leakage during the exchange of model updates between devices and the central server, ensuring data confidentiality. Nevertheless, the study noted that SA requires significant communication resources, which could negatively impact network performance and increase bandwidth consumption—an important challenge to overcome when applying federated learning in resource-constrained environments.

The study concluded that current privacy protection techniques offer effective solutions, but they still need improvements to ensure efficient resource utilization, particularly in constrained environments like IoT. The researchers recommended the development of new solutions that balance privacy, model accuracy, and energy consumption efficiency to ensure the effective and sustainable deployment of federated learning in practical applications.

This study [30] addressed the privacy challenges of Federated Learning (FL) from the perspective of compliance with the General Data Protection Regulation (GDPR). The researchers analyzed the methods used to ensure FL's compliance with these regulations while maintaining model efficiency. The study clarified that, despite being a promising alternative to centralized models in terms of security, FL does not automatically guarantee full data protection, as sensitive information can be inferred from the model updates exchanged between participants.

Among the techniques studied, differential privacy (DP) emerged as a key solution, as it reduces the risk of inferring individual data by adding noise to the model updates, requiring a balance between privacy and performance. Additionally, the role of homomorphic encryption (HE) was analyzed, allowing computations on encrypted data, thereby providing an advanced level of security. However, the high computational cost of this technique limits its application in real-world scenarios.

Furthermore, the study explored secure aggregation (SA) as a means to protect the confidentiality of updates sent between clients and the central server. Despite its effectiveness in preventing information leakage, the high bandwidth consumption may impact the system's efficiency when scaling FL to include more participants. The study noted that finding alternative solutions, such as improving encryption algorithms and adaptive noise techniques, may be necessary to address these challenges and ensure FL's compliance with GDPR requirements.

The researchers concluded that full GDPR compliance in FL environments still requires ongoing improvements, as current methods do not provide absolute protection without affecting model performance. The study recommended developing hybrid techniques that combine differential privacy and efficient encryption, with a focus on improving communication efficiency and reducing computational costs to ensure a safer and more effective FL environment for practical applications. This study [31] focused on evaluating the efficiency and privacy of Federated Learning (FL) algorithms in practical environments, with an emphasis on the trade-off between performance and data protection. The researchers analyzed the performance of the FedAvg and FedProx algorithms, two of the most commonly used methods in FL, by comparing their accuracy and efficiency in handling imbalanced data and heterogeneous distribution across participants.

The results showed that FedProx offers greater capability in handling data disparity between different devices, making it more stable in heterogeneous distribution scenarios. However, this advantage comes at the cost of increased computational consumption, which may lead to higher operational costs, especially in resource-limited environments. On the other hand, FedAvg proved to be efficient in terms of resource consumption but was more sensitive to data differences between clients, which negatively affected the model's stability in some cases.

Additionally, the study tested the impact of adding privacy techniques, such as differential privacy (DP), on model performance. The study demonstrated that using high noise

levels for protection could lead to a decrease in model accuracy, requiring precise tuning of parameters to achieve a balance between security and performance. The efficiency of secure aggregation (SA) was also examined, showing that it improves data protection but consumes significant bandwidth, which could affect communication efficiency, particularly when the number of participants in the system increases.

The researchers concluded that the efficiency of FL algorithms heavily depends on the nature and distribution of the data among clients, and improving performance requires dynamic adjustment of parameters according to the application environment. The study recommended developing adaptive algorithms that can strike a balance between resource consumption, model accuracy, and data protection to ensure efficient FL operation in real-world scenarios.

This study [35] focused on developing a verifiable federated learning (VFL) model aimed at enhancing the credibility and security of models trained in industrial Internet of Things (IIoT) environments. The researchers addressed the challenges associated with ensuring the validity of federated updates while preserving data privacy and not affecting performance efficiency.

The study proposed using Lagrange Interpolation as a method to verify the updates aggregated at the central server without revealing the original data. Experiments showed that this approach allows for detecting any tampering in updates that might occur due to internal attacks or untrusted servers, thereby enhancing the overall security of the model.

Additionally, blinding technology was employed to prevent the extraction of sensitive information from the transmitted updates, providing extra protection against gradient analysis attacks. However, the study showed that this approach could lead to increased computational overhead, especially when dealing with a large number of participants in the industrial network.

The researchers concluded that VFL provides strong protection against attacks targeting the integrity of federated models, but further improvements are needed in resource consumption and communication efficiency to ensure its applicability in large-scale industrial environments. The study recommended the development of more efficient aggregation algorithms that can balance security and computational performance to ensure the sustainable operation of FL in IIoT environments.

In this study [47], the researchers analyzed the security aspects of federated learning (FL) in the Internet of Things (IoT), identifying key threats such as inference attacks, poisoning, and eavesdropping on data during transmission. The study reviewed protection strategies, including homomorphic encryption (HE), secure aggregation (SA), and noise addition to maintain privacy.

The researchers proposed solutions such as identity-based authorization and decentralized hashing techniques to enhance data security, noting that these techniques may increase energy consumption and slow down data transmission. The study concluded that there is a need to develop more efficient FL algorithms and improve compression

techniques to ensure security without significantly impacting performance.

This study [36] focused on analyzing the security and privacy issues in federated learning (FL), highlighting the key threats that could affect the integrity and reliability of federated models. Among these threats, poisoning attacks were identified as one of the most common risks, where attackers inject harmful data during model training, intentionally skewing its outputs. Additionally, backdoor attacks, a form of poisoning, are used to plant a "backdoor" within the model so that it appears normal during regular testing but produces distorted outputs when targeted data is input by the attacker. Moreover, the study discussed GAN-based attacks, which exploit competitive AI networks to generate data capable of revealing sensitive information from the trained model, such as reconstructing the original training samples. Inference attacks also represent another threat, where attackers attempt to extract information about participants' data by analyzing model updates or outputs, even without access to the data itself.

To address these risks, the researchers reviewed several privacy protection techniques, such as Homomorphic Encryption (HE), which prevents access to sensitive data during training. However, the high computational cost of this approach makes it impractical in some low-resource environments.

Additionally, Secure Multi-Party Computation (SMPC) was emphasized as one of the prominent data protection methods in FL. SMPC distributes computational operations across multiple parties so that they can collaborate on computing a joint model without having to share their actual data. This approach ensures that each party holds only a part of the computational data, preventing any single entity from reconstructing the original data. However, the high computational complexity and low communication efficiency pose challenges in deploying SMPC at a large scale.

The study concluded that FL remains vulnerable to serious security risks, despite being a promising alternative to centralized models. The researchers recommended the development of hybrid solutions combining encryption, secure aggregation, and differential privacy to ensure a balance between performance and protection. They also emphasized the importance of enhancing verifiable federated learning (Verifiable FL) techniques, where updates can be validated without disclosing data, helping to prevent model manipulation and ensuring the integrity of federated learning.

This study [37] reviewed a new framework for federated learning (FEDF) designed to improve privacy preservation and accelerate training processes through parallel learning. This framework allows model training on geographically distributed data without exposing sensitive information, enhancing data security during the training process.

The researchers developed a new communication protocol that facilitates information exchange between the central server and connected devices, reducing the amount of data transmitted by up to 34 % compared to traditional approaches. The convergence of the model was also demonstrated during training using this framework, ensuring comparable accuracy to centrally trained models, with a speedup of up to 4.8× compared to traditional

training methods.

Despite these benefits, the study noted that achieving a balance between speed and privacy protection remains a challenge, as advanced protection techniques may increase computational complexity. The researchers recommended further improvements in managing communication between devices to ensure higher efficiency when scaling federated learning.

This study [38] introduced a new framework for federated learning (FL) based on hyperdimensional computing and differential privacy, aiming to balance performance and data protection in dynamic environments such as the Internet of Things (IoT). The researchers pointed out that traditional methods like adding random noise to preserve privacy can lead to a gradual deterioration in the performance of federated models over time, reducing their effectiveness in continuous learning scenarios.

To address this issue, the researchers proposed an approach called FedHDPrivacy, which leverages hyperdimensional computing (HDC) to enhance data security, while applying differential privacy adaptively, tracking accumulated noise and adding computed levels only when necessary. Experiments showed that FedHDPrivacy outperformed traditional FL models like FedAvg, FedProx, and FedNova by up to 38 % in terms of performance, with a reduced negative impact on model accuracy.

The study concluded that intelligent noise control can improve the efficiency of federated learning without compromising privacy. The researchers recommended expanding this approach to include multiple applications, such as multimodal data fusion, to further enhance the accuracy of federated models.

This study [39] focused on developing a deep federated learning framework (FedPC) designed to improve communication efficiency and protect privacy during distributed model training, especially in environments with bandwidth constraints. The researchers reviewed the high costs of data transmission in FL, where sharing trained models can consume significant communication resources, slowing down the federated learning process and affecting its performance.

To address this issue, the researchers developed a new communication protocol based on gradient compression, which reduces the amount of information exchanged between clients and the central server by up to 42.2% compared to traditional methods, allowing faster update processes without compromising model quality. Homomorphic encryption (HE) techniques were also integrated to protect the data during transmission, ensuring the confidentiality of information even if the central server is compromised.

The results showed that FedPC achieved a performance close to centrally trained models (within 8.5% of the central model performance) while significantly reducing communication costs, making it a promising solution for practical applications in resource-constrained environments. The researchers recommended expanding the research to include an analysis of the impact of network delay on model performance, aiming to improve federated learning efficiency in asynchronous environments.

2.2 Explainable AI and Federated Learning

The researchers in this study [29] aimed to explore the relationship between Federated Learning (FL) and Explainable Artificial Intelligence (XAI), with the goal of improving the transparency of federated models while maintaining privacy. By reviewing 37 previous studies, the researchers found that most research efforts focused on developing interpretation methods based on feature importance analysis, where the impact of each variable on the model's output is explained. However, the results showed that the decentralized nature of FL negatively affects the consistency of these interpretations, as the differences in local data for each client lead to variation in analysis outcomes between different nodes.

To overcome these challenges, secure aggregation (SA) was employed to ensure the exchange of interpretative information between participants without compromising data privacy, helping to build a deeper understanding of models trained across different devices. However, the study noted that, despite its security benefits, SA increases the complexity of communication processes and consumes significant bandwidth, which could impact the efficiency of FL, especially in resource-constrained environments.

Among the solutions explored, some researchers suggested integrating XAI techniques directly within FL algorithms, so that interpretative layers are included within the training processes without the need to reprocess data after learning. However, standardized criteria to achieve this goal effectively have not yet been defined, indicating the need for further research to develop adaptive interpretation methods for FL that ensure a balance between transparency, accuracy, and resource efficiency.

The study concluded that integrating XAI with FL is a necessary step towards enhancing trust in federated models, but it still faces fundamental challenges, particularly concerning the consistency of interpretations, improving communication efficiency, and reducing computational costs. The researchers recommended the development of standardized interpretation methodologies that align with the dynamic nature of FL environments, with a focus on resource-efficient techniques to ensure practical applicability.

This study [32] addressed the issue of building trust in Federated Learning (FL) by developing the FederatedTrust framework, which aims to evaluate the reliability of federated models and improve their transparency. The researchers focused on identifying the key pillars affecting FL reliability, including privacy, robustness, fairness, transparency, accountability, and federated collaboration, and provided over 30 indicators to measure these factors accurately.

FederatedTrust was integrated into the FederatedScope environment to test its performance using different datasets. The results showed that the proposed framework helps evaluate the quality of the model without affecting its efficiency. The study also demonstrated that relying on specific reliability metrics can improve the performance of federated models, especially in scenarios that require a balance between security and accuracy.

One of the key challenges highlighted was the need to improve computational effi-

ciency, as the researchers found that some transparency-enhancing techniques, such as Explainable AI (XAI), could lead to increased resource consumption. Furthermore, the study revealed that applying strategies to enhance fairness and balance among participants still requires improvement to ensure that no node is favored over others during the training and updating processes.

The study concluded that integrating trust metrics within FL could enhance its acceptance in critical applications such as healthcare and cybersecurity, but it requires further research to ensure a balance between transparency and efficiency. The researchers recommended the development of optimized algorithms that reduce resource consumption while maintaining the accuracy of evaluations, to ensure broader adoption of FL in real-world environments.

This thesis [40] focused on developing a framework that combines federated learning (FL) and explainable artificial intelligence (XAI) to improve security and privacy protection in smart healthcare systems. The researcher discussed how federated learning can be applied in healthcare to share models between hospitals without sharing raw data, thereby reducing the risks of leaking sensitive information.

Explainable AI models were proposed to help doctors understand the decisions made by trained models, especially in medical diagnosis and heart disorder detection through ECG signal analysis. Techniques such as lightweight federated learning were integrated to reduce computational resource consumption, making it more efficient for low-power medical devices.

The study discussed security challenges facing FL in medical systems, such as poisoning attacks, inference attacks, and Byzantine attacks, and presented solutions to protect data, such as homomorphic encryption (HE) and verifiable federated learning (Verifiable FL) to ensure the integrity of updates.

The thesis concluded that combining FL and XAI enhances the reliability of federated models, but still requires improvements in handling variations between medical data, reducing energy consumption, and improving communication efficiency to ensure broader adoption in clinical settings.

This study [43] focused on the development of an explainable federated learning framework (Fed-XAI), aimed at balancing privacy protection and the transparency of federated models. The researchers emphasized how to train XAI models in FL environments without compromising learning efficiency or the accuracy of interpretations.

The study proposed using decision trees and fuzzy logic models instead of traditional neural networks, as these models provide clear explanations without the need for post-hoc interpretation techniques. Several model aggregation strategies were compared, such as FedAvg and FedProx, with results showing that using FL with XAI can reduce the risk of information loss during training.

The study concluded that combining FL with XAI enhances the reliability of federated models, but further research is needed to develop standardized guidelines to ensure accu-

rate interpretations without impacting performance. This study [44] focused on the use of federated learning (FL) and explainable artificial intelligence (XAI) in financial fraud detection. The researchers discussed the challenges faced by traditional systems, such as data imbalance, privacy protection, and the lack of transparency in model decisions.

A deep neural network (DNN) model was developed specifically to detect fraudulent patterns in banking data, relying on FL to train the model across multiple financial institutions without sharing raw customer data. XAI techniques such as SHAP were integrated to provide clear explanations of the model's decisions, enhancing user trust.

The results showed that the proposed approach achieves high accuracy in fraud detection while maintaining data privacy and improving decision transparency. The researchers recommended expanding the research to analyze the impact of FL on interpretation performance in various financial scenarios.

2.3 Federated Learning in IoT and Healthcare

This study [33] focused on developing a new security framework to protect Internet of Medical Things (IoMT) systems using Federated Learning (FL). The proposed model relies on Artificial Neural Networks (ANN) for threat detection, incorporating mechanisms that preserve data privacy during training.

The model was enhanced with Explainable AI (XAI) to improve transparency in the decision-making process, allowing doctors and users to understand how the model arrived at its outputs. The experiments demonstrated that the combination of FL and XAI achieves performance comparable to centralized models in terms of accuracy, with the added benefit of preserving data privacy.

One of the challenges faced in the study was resource consumption. The results showed that secure aggregation (SA) techniques, used to protect updates between devices, could lead to increased bandwidth consumption, which could impact the performance of real-time medical systems. The study also pointed out that the latency in federated training processes might be a limiting factor in certain critical applications, such as real-time medical monitoring.

The researchers concluded that FL can enhance the security and privacy of IoMT systems, but further improvements are needed in resource consumption and communication efficiency. The study recommended the development of adaptive algorithms capable of balancing security and performance requirements to ensure more effective use of FL in biomedical applications.

This study [34] discussed the challenges associated with implementing Federated Learning (FL) in smart healthcare systems, focusing on how to improve the efficiency of federated models when dealing with heterogeneous data distributed across different medical institutions. Unlike studies that focused on data security, the researchers here concentrated on the operational and performance aspects that affect the effectiveness of FL in

medical applications.

One of the key challenges highlighted was the significant variation in data quality across institutions, where differences in medical devices and data collection protocols lead to issues with consistency and generalization, potentially resulting in imbalanced models. The study also pointed out that the available computational resources vary across medical centers, making the efficient implementation of FL complex, especially when there is latency in model updates.

The researchers proposed solutions such as asynchronous federated learning, which allows devices to update models according to their computational capabilities without the need for full synchronization with all participants. Communication optimization techniques and reducing resource consumption were also explored, particularly in medical environments that require rapid and efficient response, such as emergency units.

The study concluded that the application of FL in healthcare still requires significant improvements in managing data heterogeneity, enhancing communication efficiency, and ensuring fairness in training across institutions. The researchers recommended developing customized protocols to adapt FL to the specific requirements of each medical environment, enabling more efficient use of FL in future healthcare systems.

This study[42] focused on federated learning in analyzing healthcare data with an emphasis on privacy preservation. As the reliance on big medical data increases for analyzing health conditions and predicting diseases, challenges related to privacy protection and compliance with regulations such as the General Data Protection Regulation (GDPR) emerge. The researchers proposed a federated learning framework for analyzing healthcare data, which allows for preserving patient privacy while maintaining model accuracy. Various challenges were examined, including equal access to data and data distribution across different nodes. The opportunities provided by this technology include improving the efficiency of healthcare systems and reducing the risk of leakage of sensitive data. The study also suggested that differential privacy and encryption techniques could play a key role in enhancing security and privacy during federated training. The study showed that federated learning is a promising option for healthcare systems, enabling effective health models to be trained without disclosing patient data, while emphasizing the need for optimized algorithms to ensure good model performance and achieve a balance between accuracy and privacy in multi-party environments. In the future, these technologies face challenges in adapting to the vast diversity of healthcare data, as well as the need to balance security and communication efficiency in networks with limited capacity.

This study [41] explored the possibility of integrating federated learning (FL) and explainable artificial intelligence (XAI) in sixth-generation (6G) networks, with a focus on their application in intelligent vehicle systems (V2X). The researchers highlighted that XAI enhances the transparency of autonomous driving systems, helping to make more reliable decisions in vehicle and traffic management.

A Fed-XAI framework was proposed to evaluate the quality of service in smart vehicle

networks, where FL ensures data privacy during training, while XAI provides clear explanations for model decisions. The study also analyzed the benefits of federated learning in reducing communication costs and protecting user privacy, especially when dealing with sensitive data such as locations and routes.

Experiments showed that FL improves communication efficiency and reduces resource consumption compared to central models, but it still faces challenges related to the large variance in vehicle data and the impact of network latency on federated model quality. The researchers recommended the development of standardized criteria for Fed-XAI systems to enhance their adoption in industrial applications and autonomous vehicles.

2.4 Federated Learning in Advanced Systems

In this study [45], the researchers developed a model based on federated learning (FL) to combat distributed denial of service (DDoS) attacks in Internet of Things (IoT) networks. FL enables connected devices to collaborate in training a global model without sharing their sensitive data, thus reducing the risk of data leakage.

The researchers used advanced aggregation algorithms such as FedAvg and FedAvgM to improve the stability and accuracy of the model when dealing with non-IID data. Additionally, deep autoencoders were adopted to analyze and reduce the dimensionality of the data, aiming to improve performance and enhance attack detection accuracy. The results showed that FedAvgM outperformed FedAvg in handling heterogeneous data, providing higher stability in detecting attacks while reducing false positive rates.

The study concluded that FL can be an effective solution for protecting IoT networks from DDoS attacks, but it faces challenges related to increased resource consumption as the application scale expands. The researchers recommended integrating FL with other techniques such as encryption and behavior analysis to improve detection efficiency and reduce energy consumption.

In this study [46], the researchers conducted a comprehensive review of the use of federated learning (FL) supported by blockchain technology to enhance security in the Internet of Things (IoT). The study analyzed the advantages provided by this integration, such as improved privacy, reduced reliance on a central server, and enhanced trust in distributed learning models. The study discussed the key challenges facing FL in IoT, such as compatibility issues, high resource consumption, and data heterogeneity.

To achieve security, the study proposed the use of consensus mechanisms in blockchain to ensure the reliability of model updates and prevent tampering. However, the study pointed out that these mechanisms impose high computational costs, which may limit their use in resource-constrained devices. Additionally, the need to develop hybrid solutions combining FL and blockchain more efficiently to reduce complexity and improve performance was highlighted.

The study concluded that combining FL and blockchain provides a high level of secu-

rity, but still faces challenges in scalability and effective use in IoT environments. The researchers recommended improving federated learning algorithms, using data compression techniques, and developing more adaptive models for resource-constrained environments to achieve a balance between security and efficiency.

Title	**Authors and Institution**	**Year**	**Focus Areas**	**Limitations**
Privacy Threats and Countermeasures in Federated Learning for IoT	Adel ElZemity, Budi Arief (University of Kent)	2024	Comprehensive analysis of privacy risks and proposed countermeasures in IoT-based FL environments.	Implementation complexity across heterogeneous IoT devices.
Interplay between FL and XAI: A Scoping Review	Luis M. Lopez-Ramos et al. (VALIDATE Consortium)	2024	Studied the integration of FL and XAI in various sectors.	Absence of standard evaluation metrics for interpretability.
Privacy Preservation in FL from a GDPR Perspective	Nguyen Truong et al. (Imperial College London)	2021	Surveyed FL compliance with GDPR focusing on legal and technical dimensions.	Limited insights into practical deployment issues.
Empirical Study of FL Efficiency and Privacy	Sofia Zahri et al. (QMUL, TU-Dublin)	2024	Benchmarking of FL algorithms in terms of privacy and training efficiency.	Evaluation on large-scale and diverse datasets was lacking.
FederatedTrust: A Solution for Trustworthy FL	Pedro M. Sánchez Sánchez et al. (University of Murcia, Zurich)	2023	Proposed a framework using trust scores and reputation mechanisms in FL.	High computation overhead not suitable for edge devices.
hExplainable ML Security Framework for IoMT	Si-ahmed Ayoub et al. (Blida 1 University, Emory University)	2024	Developed XAI-based framework for privacy in medical IoT systems.	Demands high computational resources in practical deployments.
FL-Based AI in Smart Healthcare: Challenges	Anichur Rahman et al. (KSU, Aalto Univ.)	2023	Offered taxonomy of FL-based AI solutions in healthcare and highlighted open issues.	Lack of real-time systems and energy-efficient models.
VFL: Verifiable FL for Industrial IoT	Anmin Fu et al. (Nanjing Univ. of Science and Tech.)	2020	Introduced verifiable FL model for industrial IoT using privacy-preserving techniques.	Verification mechanism adds delay and reduces performance.
Privacy and Security in Federated Learning	Rémi Gosselin et al. (Appl. Sci. Journal)	2022	Surveyed privacy risks and security solutions in FL models.	Did not consider impact of integration with 6G networks.
Privacy-preserving and Parallel FL Framework	Tien-Dung Cao et al. (Tan Tao University)	2021	Proposed an efficient FL framework with privacy guarantees and parallel processing.	Communication cost under mobility not fully analyzed.

Title	Authors and Institution	Year	Focus Areas	Limitations
FL for Communication Efficiency and Privacy	Tien-Dung Cao et al. (SIT, Vietnam Nat'l Univ.)	2022	Built a FL model to balance communication efficiency and data privacy.	Lacks real-time system deployment and validation.
FL with Differentially Private Hyperdimensional Computing	Fardin Jalil Piran et al. (UC Irvine)	2024	Combined FL with HDC and differential privacy to process high-dimensional data.	Scalability issues due to computational complexity.
Federated Learning for Privacy-Preserving Healthcare Analytics	S. S. Pradeep et al. (Int. J. Healthcare Inf. Syst. Informatics)	2023	Proposed a federated learning framework for privacy-preserving healthcare analytics, focusing on improving patient data privacy and ensuring model efficiency.	Managing heterogeneous healthcare data, improving communication efficiency, and balancing security with system performance in medical environments.
FL with XAI for Smart Healthcare	Ali Raza (Univ. of Lille & Kent)	2023	Integrated FL and XAI for interpretable models in healthcare diagnostics.	Data heterogeneity affects accuracy and training stability.
FL of Explainable AI Models in 6G	Alessandro Renda et al. (Information Journal)	2022	Developed explainable FL models for future 6G systems in autonomous vehicles.	Latency and network performance constraints still exist.
Fed-XAI: Federated Learning of XAI Models	José L. Corcuera Bárcena et al. (Univ. of Pisa)	2022	Built a FL framework using decision trees and fuzzy logic for interpretability.	Standardization needed in explanation outputs.
Transparency and Privacy in Financial Fraud Detection	Tomisin Awosika et al. (Anglia Ruskin Univ.)	2023	Merged FL and SHAP for interpretable financial fraud detection.	Edge device interpretation accuracy not extensively tested.
FL Against DDoS Attacks in IoT	Ghazaleh Shirvani et al. (Preprint arXiv)	2024	Applied deep FL with autoencoders to mitigate DDoS attacks in IoT.	Higher energy consumption with scale-out networks.
Blockchained Federated Learning for IoT	Yanna Jiang et al. (ACM Comp. Surveys)	2024	Merged blockchain and FL to create secure decentralized IoT networks.	High cost due to blockchain consensus protocols.
Blockchain-based FL for IoT Security	Wael Issa et al. (ACM Comp. Surveys)	2023	Surveyed integration of blockchain and FL for secure IoT systems.	Scalability limitations of blockchain remain unresolved.

Table 2.1: Summary of Selected Studies 1–20

CHAPTER 3

Chapter 3

Proposed method

Federated Learning (FL) has emerged as a key solution to address privacy and data security challenges in healthcare applications. By allowing machine learning models to be trained locally on edge devices—such as hospital systems—without sharing sensitive patient data, FL preserves confidentiality and reduces communication overhead.

While Internet of Medical Things (IoMT) systems connect medical devices to enable continuous patient monitoring, this study does not propose a novel IoMT architecture. Instead, it focuses on the integration of federated learning into existing healthcare infrastructures to enhance data privacy, scalability, and model efficiency.

This chapter outlines the general framework of the proposed method, which seeks to strike a balance between privacy preservation and model interpretability in a federated environment. The approach consists of several stages: starting with secure data collection from hospitals, progressing through local model training on individual clients, and culminating in centralized aggregation of model updates. The final stage involves the interpretation of results using visual explanation techniques such as GradCAM and NormGrad.

By embedding interpretability mechanisms within the FL workflow—without compromising data privacy—the proposed framework promotes transparency and fosters trust in AI-driven decisions, particularly in critical domains like healthcare and IoT-based systems.

3.1 The Overall Framework of the Proposed Approach

The proposed system in this study is based on an **interpretable federated learning framework** designed to enhance privacy protection within healthcare environments. This system operates in a collaborative setting among multiple medical institutions (such as hospitals), where sensitive health data is retained locally within each hospital. Information exchange is limited exclusively to model updates, without any direct sharing of raw data.

The design focuses on establishing an integrated architectural framework comprising the following stages:

- Local data collection within each hospital.
- Local model training using patient data available at each site.
- Interpretation of model outputs using explainable AI techniques (GradCAM and NormGrad) at each hospital to clarify the reasoning behind predictions.
- Centralized model aggregation through a main server responsible for merging parameters received from clients.

The following figure illustrates the overall architectural overview of the system:

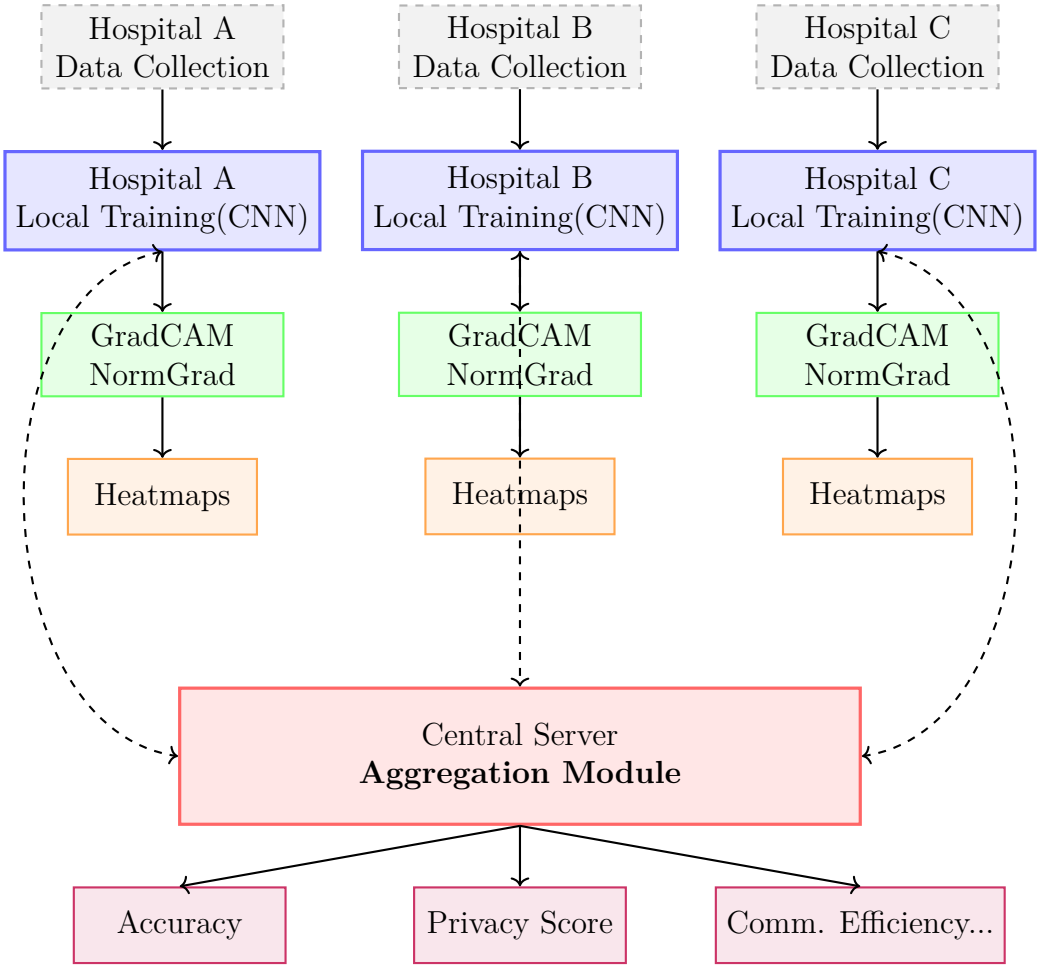

Figure 3.1: Interpretable Federated Learning Architecture for Healthcare

To delve deeper, the following sections will explain each stage of the system in detail:

3.1.1 Data Collection Stage

The data is collected in hospital settings using professional medical scanners, ensuring high accuracy and quality that enhance the reliability of results and analysis.

- Given the sensitive and private nature of this data, it is handled from the outset with strict security protocols.

- The collected information typically includes vital signs and medical images, which can later be processed and analyzed for various diagnostic and research purposes.

3.1.2 Local Model Training Stage

An efficient Convolutional Neural Network (CNN) model was developed due to its high capability in processing image data and extracting spatial patterns. The proposed model consists of the following architecture:

- Two consecutive **convolutional layers**, used to extract fundamental features from images such as edges and shapes, each followed by a **max pooling layer** to reduce dimensionality and computational complexity.
- Three **fully connected layers**, used to integrate the extracted features from the previous layers and perform the final classification over the ten classes in the CIFAR-10 dataset.
- The **CrossEntropy Loss** function was used, as it is well-suited for multi-class classification tasks, measuring the divergence between the model's predicted distribution and the true class distribution.

$$\mathcal{L}_{\text{CrossEntropy}} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \tag{3.1}$$

$$\text{where: } \begin{cases} C & \text{is the number of classes,} \\ y_i & \text{is the true label for class } i \text{ (0 or 1),} \\ \hat{y}_i & \text{is the predicted probability for class } i. \end{cases}$$

For a batch of N samples:

$$\mathcal{L} = -\frac{1}{N} \sum_{n=1}^{N} \sum_{i=1}^{C} y_i^{(n)} \log(\hat{y}_i^{(n)}) \tag{3.2}$$

- The **Adam Optimizer** was adopted for its fast convergence and stable parameter updates. It combines the benefits of SGD (which updates parameters based on the average gradient from random mini-batches) and RMSProp (which adaptively adjusts the learning rate for each parameter based on the moving average of squared gradients).

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$$

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}$$

$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t}$$

$$\theta_{t+1} = \theta_t - \alpha \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \tag{3.3}$$

where:

$$\begin{cases} m_t & \text{is the first moment estimate (mean of gradients),} \\ v_t & \text{is the second moment estimate (uncentered variance of gradients),} \\ \hat{m}_t, \hat{v}_t & \text{are bias-corrected moment estimates,} \\ \theta_t & \text{are the parameters at iteration } t, \\ \alpha & \text{is the learning rate,} \\ \beta_1, \beta_2 & \text{are exponential decay rates for the moment estimates,} \\ g_t & \text{is the gradient at iteration } t, \\ \epsilon & \text{is a small constant to avoid division by zero.} \end{cases}$$

This architecture strikes a balance between computational efficiency and accuracy, making it suitable for deployment in federated learning (FL) environments with non-IID data distributed across multiple clients.

3.1.3 Interpretability Mechanisms

After each training round, the model is analyzed using two interpretive techniques,To enhance the understanding of the model's behavior and analyze its decisions,offering deep insights into how inputs influence the model's outputs:

1. **GradCAM (Gradient-weighted Class Activation Mapping)**
   GradCAM is an advanced technique for analyzing the model's classification responses. It utilizes the gradients computed during the backpropagation process to identify important regions in the image that significantly contribute to the model's decision. This approach allows researchers to visually interpret the regions that the model focuses on, providing a better understanding of how it processes data and makes predictions[48].

The mathematical formulation of GradCAM involves the following steps:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{3.4}$$

$$L_{\text{GradCAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k \right) \tag{3.5}$$

where:

- y^c is the score for class c (before the softmax),
- A^k is the k-th feature map of the convolutional layer,
- A_{ij}^k is the activation at spatial location (i, j) in feature map k,
- Z is the number of pixels in the feature map (i.e., $Z = \sum_i \sum_j 1$),
- α_k^c is the importance weight for feature map k with respect to class c,
- L_{GradCAM}^c is the GradCAM heatmap for class c,
- ReLU ensures only positive influences are considered.

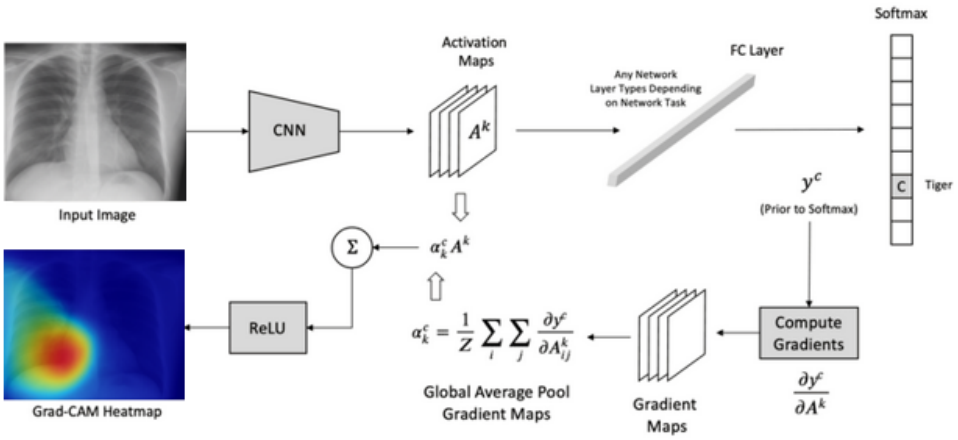

Figure 3.2: GradCAM method

2. **NormGrad (Normalized Gradient)**
   NormGrad offers an innovative method to explore the influence of absolute gradients in the image on the model's output. By analyzing the individual gradients of each pixel in the image, this technique provides a quantitative measure of each input component's impact on the final decision. This insight leads to a deeper understanding of the model's internal interactions and contributes to transparency in the computations performed by the neural network[49].

   The mathematical formulation of NormGrad can be expressed as:

$$\hat{H}(x) = \frac{\|\nabla_x f_y(x)\| - \min\left(\|\nabla_x f_y(x)\|\right)}{\max\left(\|\nabla_x f_y(x)\|\right) - \min\left(\|\nabla_x f_y(x)\|\right) + \varepsilon} \tag{3.6}$$

Where:

- $\hat{H}(x)$ is the normalized saliency heatmap of input x.
- $f_y(x)$ is the model's output score for the predicted class y.
- $\nabla_x f_y(x)$ denotes the gradient of the output score f_y with respect to the input x.
- $\| \cdot \|$ is the norm applied to the gradient (typically L1 or L2 norm).
- $\min(\cdot)$ and $\max(\cdot)$ are the minimum and maximum values over the gradient norm map.
- ε is a small constant added for numerical stability to avoid division by zero.

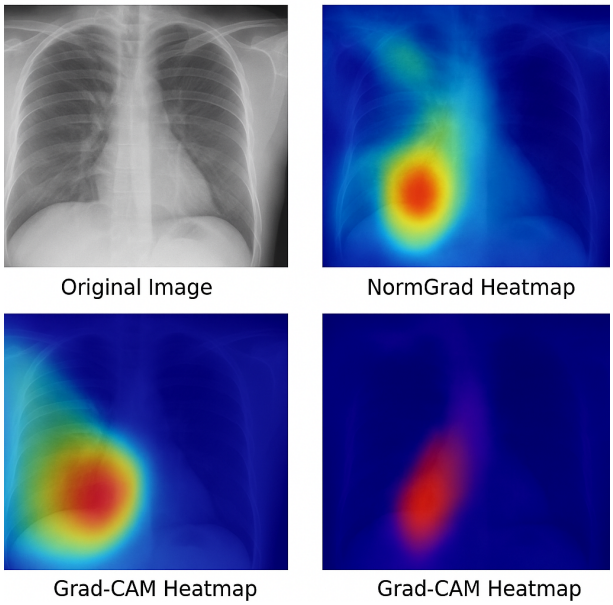

Figure 3.3: NormGrad and GradCAM Process

3.1.4 Central Aggregation Phase (Aggregation Server)

After each client completes the local training, it sends the updates (new model weights) to the central server.

- No sensitive data or images are transmitted, which enhances privacy protection.
- The communication is secured using secure protocols for data transmission.

At the central server, updates from all clients are received and aggregated using one of the following federated learning strategies:

FedAvg (Federated Averaging)

FedAvg is one of the simplest and most widely used federated learning algorithms. It relies on computing a weighted average of the model weights sent by the clients. After local training, each client sends its model weights to the central server, which then calculates the weighted average based on the number of local training samples for each client, and updates the global model.

This strategy is most effective in balanced data distribution environments and is usually applied in synchronous settings where the server waits to receive updates from all selected clients before performing the aggregation step.

$$w_{t+1} = \sum_{k=1}^{K} \frac{n_k}{n} w_t^{(k)} \tag{3.7}$$

where:

- w_{t+1} is the global model after update at round $t+1$,
- K is total number of clients,
- n_k is number of training samples at client k,
- $n = \sum_{k=1}^{K} n_k$ is total number of samples across all clients,
- $w_t^{(k)}$is model weights sent from client k after local training at round t.

FedAdam

FedAdam is an improvement of the popular Adam algorithm, adapted for federated learning environments. It uses an adaptive learning rate and momentum in the updates to accelerate convergence and provide greater stability during training, especially when dealing with heterogeneous data. FedAdam works well with both balanced and unbalanced data distributions, and is typically implemented in synchronous environments.

$$\begin{aligned} g_t &= \frac{1}{K} \sum_{k=1}^{K} \left(w_t - w_t^{(k)} \right) \\ m_t &= \beta_1 m_{t-1} + (1-\beta_1) g_t \\ v_t &= \beta_2 v_{t-1} + (1-\beta_2) g_t^2 \\ \hat{m}_t &= \frac{m_t}{1-\beta_1^t} \\ \hat{v}_t &= \frac{v_t}{1-\beta_2^t} \\ w_{t+1} &= w_t - \eta \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \end{aligned} \tag{3.8}$$

where:

- g_t is aggregated gradient at round t,
- m_t is first moment estimate (momentum),
- v_t is second moment estimate,
- β_1, β_2 is exponential decay rates for the moment estimates,
- $\hat{m}_t$, $\hat{v}_t$ is bias-corrected moment estimates,
- η is server learning rate,
- ϵ is small constant to prevent division by zero,
- w_t and w_{t+1} is global model weights at rounds t and $t+1$, respectively,
- $w_t^{(k)}$ is local model weights from client k after training.

FedYogi

FedYogi is an improvement over FedAdam that addresses oscillation issues in model updates by using a mechanism to stabilize the average of gradients. This algorithm helps stabilize the model when data is heterogeneous, outperforming FedAdam in scenarios requiring higher precision. It is usually used with unbalanced data distributions and in synchronous environments.

$$\begin{aligned} m_t &= \beta_1 m_{t-1} + (1-\beta_1) g_t \\ v_t &= v_{t-1} - (1-\beta_2) \cdot \text{sign}(v_{t-1} - g_t^2) \cdot g_t^2 \\ \hat{m}_t &= \frac{m_t}{1-\beta_1^t} \\ \hat{v}_t &= \frac{v_t}{1-\beta_2^t} \\ w_{t+1} &= w_t - \eta \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \end{aligned} \tag{3.9}$$

where:

- g_t is aggregated gradient at round t,
- m_t is first moment estimate (momentum),
- v_t is second moment estimate with Yogi correction,
- β_1, β_2 is exponential decay rates for the moment estimates,
- η is learning rate,
- ϵ is small constant to prevent division by zero,
- w_t and w_{t+1} is global model weights at rounds t and $t+1$, respectively.

FedOpt

FedOpt is a general framework that combines FedAvg with advanced optimizers such as RMSProp, Adam, or Yogi, aiming to speed up convergence and achieve more precise updates of the federated model. Due to its flexibility, it can be used in various scenarios with random data distributions and operates efficiently in synchronous environments ensuring unified aggregation from all clients.

$$\begin{cases} g_t = \frac{1}{K} \sum_{k=1}^{K} \left(w_t - w_t^{(k)} \right) \\ m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \\ v_t = v_{t-1} + \Delta v_t(g_t, v_{t-1}, \beta_2) \\ \hat{m}_t = \frac{m_t}{1 - \beta_1^t} \\ \hat{v}_t = \frac{v_t}{1 - \beta_2^t} \\ w_{t+1} = w_t - \eta \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \end{cases} \tag{3.10}$$

where:

- w_t is global model weights at round t,
- $w_t^{(k)}$ is local model weights from client k after local training,
- K is total number of clients,
- g_t is aggregated pseudo-gradient at round t,
- m_t is first moment estimate (momentum),
- v_t is second moment estimate updated by Δv_t depending on the optimizer,
- β_1, β_2 is exponential decay rates for moment estimates,
- $\hat{m}_t$, $\hat{v}_t$ is bias-corrected moment estimates,
- η is global learning rate,
- ϵ is small constant to avoid division by zero,
- $\Delta v_t(g_t, v_{t-1}, \beta_2)$ is update rule for the second moment, e.g.:
  - For **FedAdam** is $\Delta v_t = (1 - \beta_2)(g_t^2 - v_{t-1})$,
  - For **FedYogi** is $\Delta v_t = -(1 - \beta_2) \cdot \text{sign}(v_{t-1} - g_t^2) \cdot g_t^2$,
  - For other optimizers, Δv_t may vary.

The updated global model is sent to all client devices to start a new training round, ensuring continuous updating of the model at each participating node.

3.1.5 Prediction and Interpretation

At the final stages of the federated learning process, the topic of model interpretability gains significant importance, as it constitutes a fundamental element for enhancing the understanding of how models make decisions and ensuring transparency in operations. Since each client trains the model locally on its own data, the interpretation process is also performed at the client level, which ensures data privacy and protects against leakage.

Each client generates detailed explanations of the model using advanced techniques such as Grad-CAM and NormGrad, highlighting the features and information that have the greatest impact on prediction. Subsequently, only the model weights are sent to the central server, where these weights are aggregated to form a global model representing the collective knowledge of all clients.

This integrated approach combines the power of distributed learning with providing precise local interpretability for each client, while preserving data confidentiality, supporting the construction of highly reliable and interpretable federated learning models that facilitate their adoption in sensitive domains such as healthcare and the Internet of Things.

CHAPTER 4

Chapter 4

Experimental evaluation and discussion

4.1 Data Preparation

The CIFAR-10 dataset was adopted in this study for simulation and evaluation of the federated learning system's performance. This dataset contains 60,000 color images sized 32×32 pixels, distributed across 10 distinct classes, with 50,000 images allocated for training and 10,000 for testing.

As a preliminary step, the data underwent preprocessing, which included:

- **Normalization** using the mean and standard deviation computed from the training set, in order to standardize the pixel value ranges and enhance training stability.

Figure 4.1: Photo of data

After preprocessing, the data was partitioned among clients based on various distribution strategies, aimed at investigating the effects of data heterogeneity and training synchronization on the performance of federated learning. These strategies include:

- **Equal-Sync Distribution**: Data is equally distributed among clients, and all clients participate synchronously in every training round.
- **Unequal-Sync Distribution**: Clients receive varying amounts of data, while maintaining synchronous updates.
- **Random-Async Distribution**: Data is randomly and unevenly distributed, and clients participate asynchronously in the training process.

These diverse configurations were adopted to simulate realistic scenarios and reflect the true challenges present in Internet of Medical Things (IoMT) networks, where clients differ significantly in data volume and computational capabilities.

4.2 Evaluation Metrics

The performance of the federated learning system was evaluated using a comprehensive set of metrics that consider aspects of accuracy, effectiveness, and privacy, as follows:

- **Accuracy**
  The weighted average classification accuracy was calculated based on the number of samples per client, reflecting the ability of the global model to generalize across different local distributions.
- **Loss**
  The weighted average loss was adopted to provide a more accurate representation of model quality, accounting for the varying data sizes across clients.
- **Communication Efficiency**
  Measured by the total amount of data exchanged (in bytes) during training. It reflects network resource usage, a critical factor in bandwidth-constrained environments.
- **Message Exchange**
  Represents the total number of model exchanges between the server and clients per round, indicating communication load and energy cost.
- **Round Time**
  Defined as the time (in seconds) to complete one training round — from sending the model to clients to receiving and aggregating their updates. It indicates system execution efficiency.
- **Privacy Score**
  This metric indicates how well the model balances predictive performance with the preservation of data privacy, highlighting the potential risk of information leakage as accuracy increases.

Additionally, the following metrics were considered:

- **Number of Rounds** Refers to the total number of federated training rounds performed, each consisting of local training at clients followed by central aggregation of models.
- **Number of Clients** Represents the number of parties involved in federated training, participating either synchronously or asynchronously, depending on the strategy type.

4.3 A Comparative Study of Federated Learning Strategies (FedAvg, FedAdam, FedYogi, FedOpt)

This experiment aims to conduct a **comparative analysis of the performance of four federated learning strategies**: *FedAvg*, *FedAdam*, *FedYogi*, and *FedOpt*, within a **homogeneous federated environment** where data is evenly distributed across all clients in terms of size and features.

The experiment involves running each strategy over **10 training rounds**, recording the *global model accuracy* at each round, as well as the *training and evaluation accuracy of each client individually*. The setup uses a CNN model and a balanced distribution of the CIFAR-10 dataset among clients. The goal is to evaluate each strategy's ability to:

- Achieve high global accuracy.
- Ensure consistent and balanced client performance.
- Handle data characteristics effectively even in a homogeneous setting.

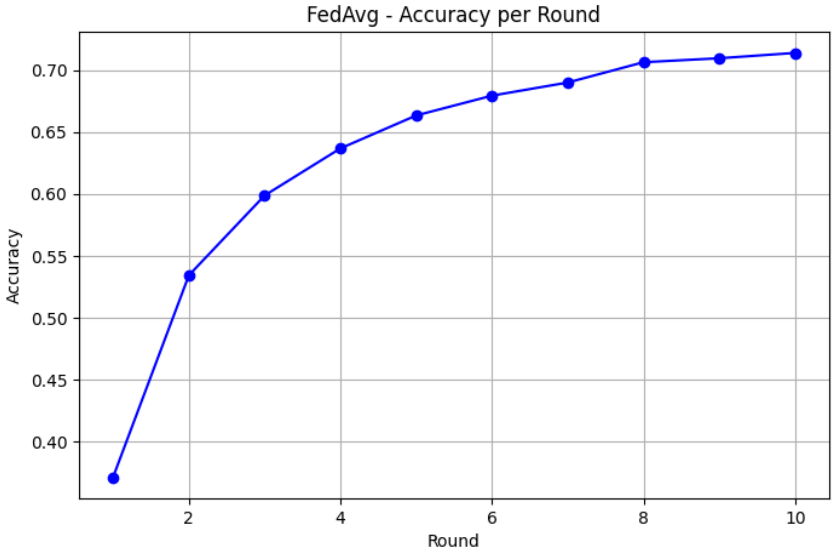

Figure 4.2: accuracy plot of FedAvg

Client Accuracy Summary Table:

Client ID	Avg Train Accuracy	Avg Eval Accuracy
0	0.7515	0.6305
1	0.7549	0.6305
2	0.7437	0.6305
3	0.7617	0.6305
4	0.7505	0.6305
5	0.7497	0.6305
6	0.7454	0.6305
7	0.7531	0.6305
8	0.7493	0.6305
9	0.7486	0.6305

Figure 4.3: local accuracy of client in FedAvg

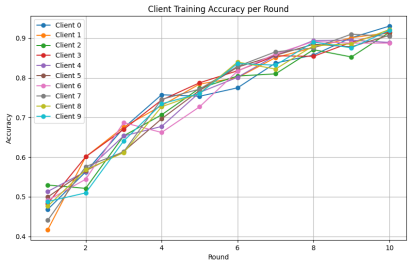

Figure 4.4: plot local accuracy of client in FedAvg

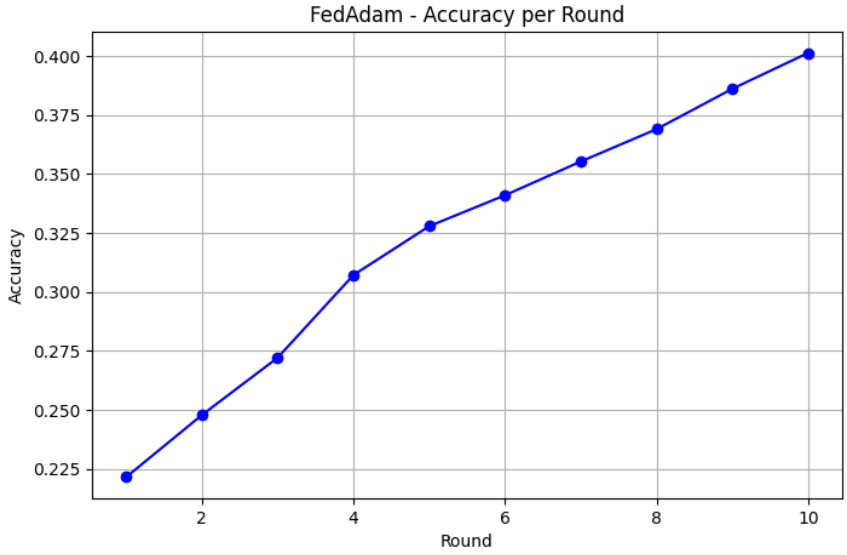

Figure 4.5: accuracy plot of FedAdam

Client Accuracy Summary Table:

Client ID	Avg Train Accuracy	Avg Eval Accuracy
0	0.4998	0.3229
1	0.5182	0.3229
2	0.5141	0.3229
3	0.5165	0.3229
4	0.5100	0.3229
5	0.5123	0.3229
6	0.5106	0.3229
7	0.5108	0.3229
8	0.5164	0.3229
9	0.5059	0.3229

Figure 4.6: local accuracy of client in FedAdam

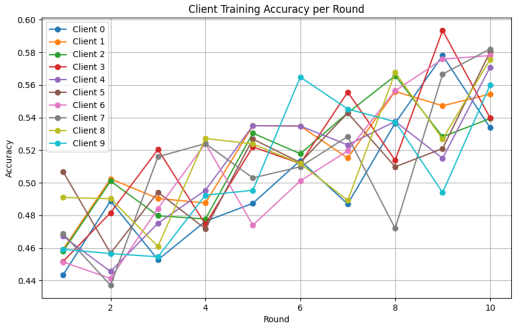

Figure 4.7: plot local accuracy of client in FedAdam

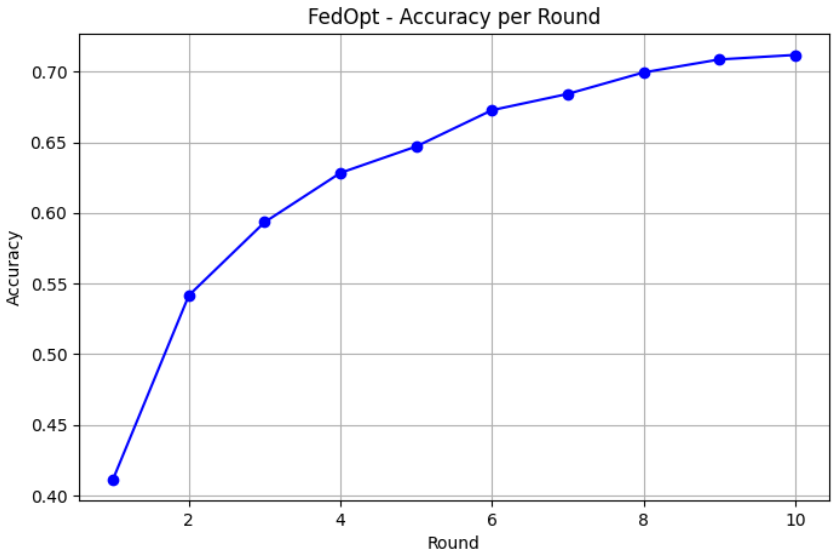

Figure 4.8: accuracy plot of FedOpt

Client Accuracy Summary Table:

Client ID	Avg Train Accuracy	Avg Eval Accuracy
0	0.7446	0.63
1	0.7425	0.63
2	0.7354	0.63
3	0.7435	0.63
4	0.7424	0.63
5	0.7531	0.63
6	0.7378	0.63
7	0.7393	0.63
8	0.7418	0.63
9	0.7531	0.63

Figure 4.9: local accuracy of client in FedOpt

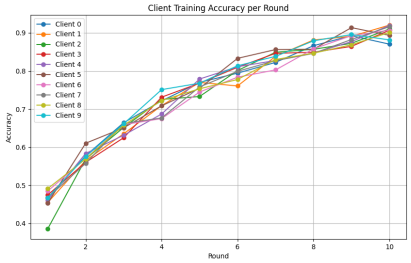

Figure 4.10: plot local accuracy of client in FedOpt

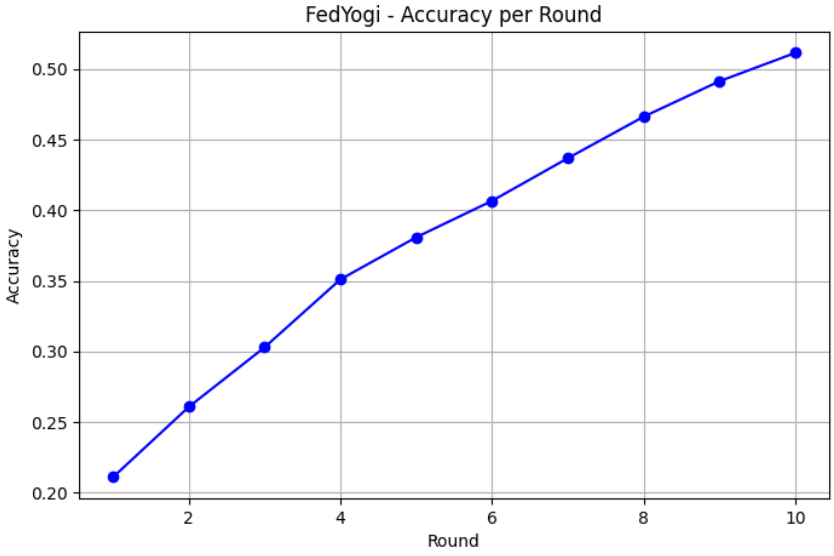

Figure 4.11: accuracy plot of FedYogi

Client Accuracy Summary Table:

Client ID	Avg Train Accuracy	Avg Eval Accuracy
0	0.5481	0.382
1	0.5549	0.382
2	0.5488	0.382
3	0.5594	0.382
4	0.5596	0.382
5	0.5584	0.382
6	0.5617	0.382
7	0.5505	0.382
8	0.5502	0.382
9	0.5558	0.382

Figure 4.12: local accuracy of client in FedYogi

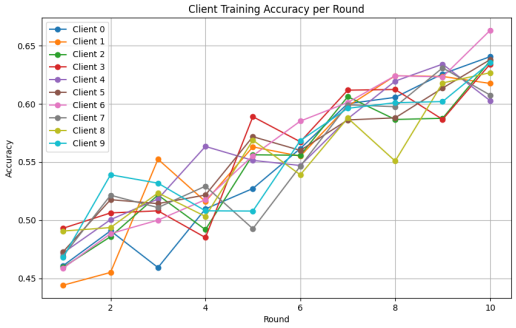

Figure 4.13: plot local accuracy of client in FedYogi

Strategy	Acc. R1	Acc. R10	Train/Eval Acc.	Performance Notes
FedAvg	0.32	0.74	0.74–0.75 / ∼0.6385	Strong and stable performance with balanced generalization. Minor variation across clients; effective in homogeneous settings.
FedOpt	0.42	0.71	0.73–0.75 / ∼0.63	Similar to FedAvg . Maintains consistent client accuracy and stable learning dynamics.
FedAdam	0.23	0.40	0.49–0.51 / ∼0.32	Lower evaluation accuracy. Significant variation across clients.
FedYogi	0.20	0.50	0.54–0.56 / ∼0.38	Limited generalization, less stable than FedAvg/FedOpt. Slightly more variance in client performance.

Table 4.1: Detailed performance summary of federated strategies with training and evaluation accuracy

Although data is distributed homogeneously, the experiment revealed **performance variability among clients** that depends on the chosen strategy.

Strategies like *FedAvg* and *FedOpt*, which rely on averaging client models, tend to **minimize variability** across clients, making them suitable for achieving overall balanced performance without fine-grained personalization.

In contrast, adaptive gradient-based strategies like *FedAdam* are **more sensitive to local data characteristics**, causing **greater variability in client performance**, even in a balanced data setting.

FedYogi shows moderate behavior; originally designed for heterogeneous environments, it does not fully benefit from the balanced distribution in this experiment.

Although there appears to be a discrepancy between the clients' accuracy results in the graph and the summary table, this discrepancy is not indicative of a programming or technical error in the implementation of the FedAvg algorithm. Rather, it is attributed to the differing nature of data representation between the two sources. The graph illustrates the progression of each client's training accuracy over multiple rounds, offering a dynamic view of the learning process, while the table presents final average values, which serve as a quantitative summary of performance.

4.4 Analysis of Data Distribution Impact on FedAvg Performance

The objective of this experiment is to analyze the effect of different data distribution patterns among clients on the performance of the **FedAvg** algorithm in a federated learning setup. Data balance is a critical factor that directly influences the accuracy and convergence stability of the global model. In cases of data imbalance, the central model may experience degradation or convergence delays, which undermines the overall effectiveness of federated learning.

To achieve this, three experimental environments were created, differing only in data distribution, while keeping all other variables constant (e.g., number of clients, number of rounds, model architecture, synchronous aggregation, and the FedAvg algorithm). The following distribution types were used:

- **Equal Sync** Each client receives an equal number of samples with a balanced representation of all classes.
- **Random Sync** Data is distributed randomly, without ensuring balance in class representation or sample size.
- **Unequal Sync** Clients receive unequal quantities of data, possibly skewed towards specific classes.

Each client's training accuracy was recorded across 10 rounds. Three separate graphs are plotted (accuracy graph for each client according to the distribution), as well as a comparative graph for overall accuracy.

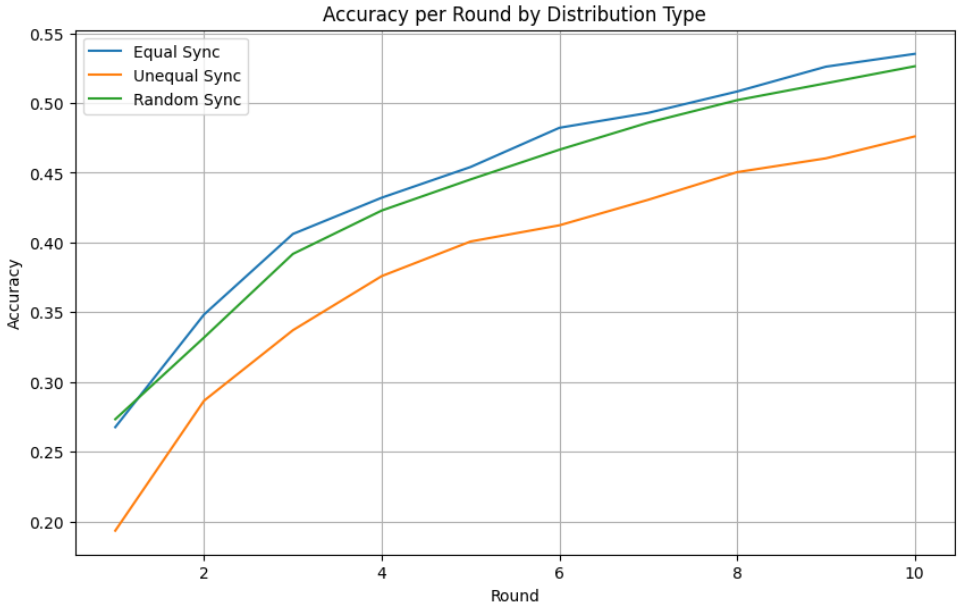

Figure 4.14: Accuracy of Equal,Random and Unequal Sync in the FedAvg

Client ID	local Accurac	global Accuracy
0	0.7467	0.6288
1	0.7443	0.6288
2	0.7517	0.6288
3	0.7587	0.6288
4	0.7402	0.6288
5	0.7545	0.6288
6	0.7546	0.6288
7	0.7473	0.6288
8	0.7417	0.6288
9	0.7374	0.6288

Figure 4.15: client accuracy in equal sync

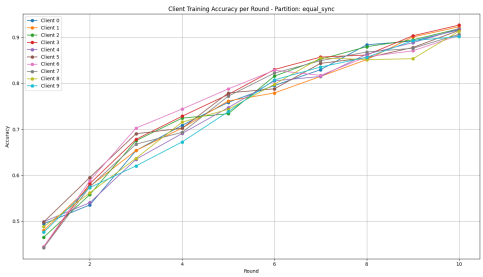

Figure 4.16: plot client accuracy in equal sync

Client ID	local Accurac	global Accuracy
0	0.7459	0.6171
1	0.7413	0.6171
2	0.7423	0.6171
3	0.7425	0.6171
4	0.746	0.6171
5	0.7446	0.6171
6	0.7461	0.6171
7	0.7418	0.6171
8	0.7341	0.6171
9	0.7341	0.6171

Figure 4.17: client accuracy in random sync

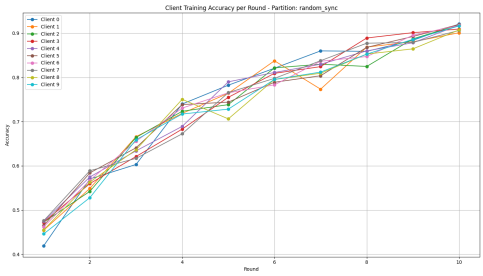

Figure 4.18: plot client accuracy in random sync

Client ID	local Accurac	global Accuracy
0	0.7288	0.5633
1	0.7317	0.5633
2	0.7107	0.5633
3	0.7271	0.5633
4	0.7019	0.5633
5	0.7081	0.5633
6	0.7103	0.5633
7	0.7222	0.5633
8	0.7226	0.5633
9	0.7217	0.5633

Figure 4.19: client accuracy in unequal sync

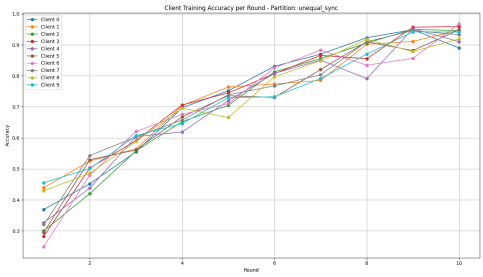

Figure 4.20: plot client accuracy in unequal sync

Table 4.2: Comparison of FedAvg Performance Across Different Data Distribution Types

Criterion	**Equal Distribution (Equal Sync)**	**Random Distribution (Random Sync)**	**Unequal Distribution (Unequal Sync)**
Central Model Accuracy	Good – reaches **0.54** by round 10 with steady improvement.	Generally good – peaks near **0.52** at best, but improves at a slower pace.	Poor performance **0.46**, with slight fluctuations.
Client Accuracy	Very consistent – most clients achieve accuracy between **0.73 and 0.75**.	good – most clients score between **0.73 and 0.74**, with some variation.	most clients are between **0.70 and 0.73**, with a few lower outliers.
Client Consistency	High – performance gaps between clients are minimal and stable.	Moderate – some crossing points between clients, but differences are not large.	The gaps between client are larger and more volatile.
Overall Distribution Quality	**Best performance** – excellent stability, consistent client results.	**Acceptable performance** –good improvement, but there's a slight drop in client alignment.	**Fair performance** – some instability and client divergence.

The results clearly indicate that the **type of data distribution** among clients significantly affects the performance of the **FedAvg** algorithm. Key scientific findings are summarized as follows:

- **Equal Sync:** Yields the best results with faster convergence, more accurate and stable global model, and minimal risk of bias or performance gaps.

- **Random Sync:** Shows moderate performance with visible variation across clients and less stable learning curves.
- **Unequal Sync:** The most problematic; leads to biased global models favoring dominant clients and produces large performance gaps among participants.

To achieve efficient federated learning using FedAvg, a balanced data distribution is essential. This is especially critical in sensitive domains like healthcare or IoT, where biased or unstable learning may result in unreliable or unsafe decisions.

4.5 The effect of Sync vs. Async on the FedAvg algorithm

An experiment was conducted to investigate the impact of synchronization modes, namely **Synchronous (Sync)** and **Asynchronous (Async)**, on the performance of the **FedAvg** algorithm within a federated learning environment using classified data.

The primary objective of this study is to evaluate how the choice of synchronization mode affects the convergence speed, final model accuracy, and stability of the global model throughout the training rounds, as well as to identify which mode achieves superior performance in terms of model quality and training efficiency.

- In **Synchronous (Sync)** mode, the server aggregates updates from all clients in each training round and updates the global model only after receiving all client updates, ensuring consistent and comprehensive model updates.
- In **Asynchronous (Async)** mode, the server updates the global model immediately upon receiving updates from any available client, allowing faster updates but potentially using incomplete or stale data, which may affect model stability and accuracy.

Furthermore, the training accuracy of each client was monitored and analyzed independently over the training rounds. This enabled the evaluation of each client's contribution to improving the global model and provided insights into the performance variability among clients, especially between the synchronous and asynchronous modes.

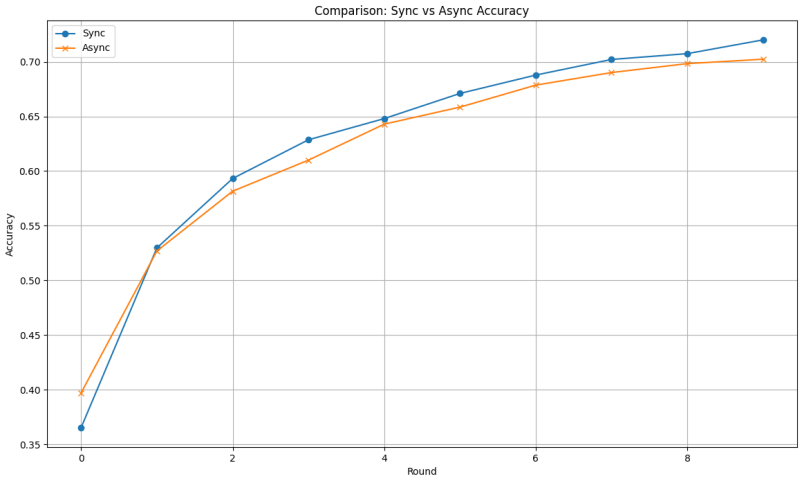

Figure 4.21: Accuracy of Sync and Async in the FedAvg

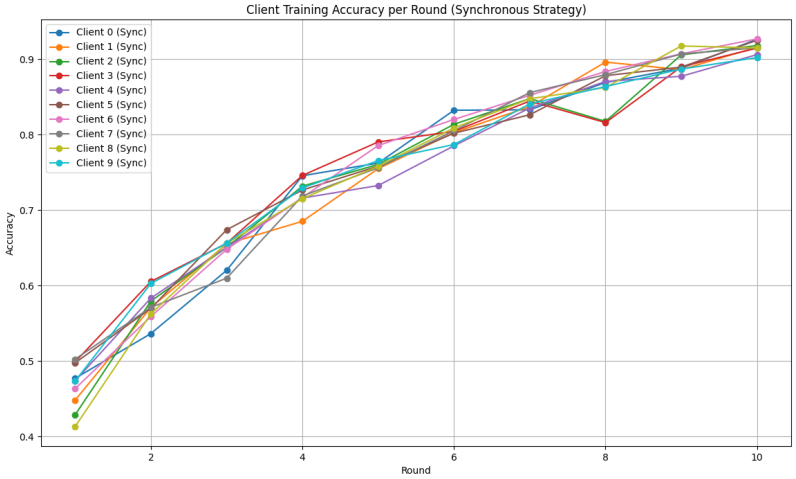

Figure 4.22: client accuracy in sync

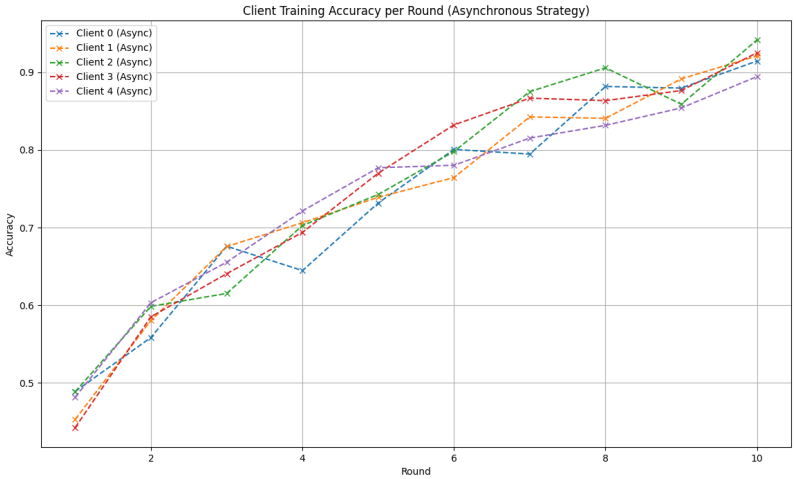

Figure 4.23: client accuracy in async

Client Accuracy Summary Table:

Client ID	local Accuracy(Sync)	local Accuracy(Async)
0	0.7491	0.7370
1	0.7453	0.7414
2	0.7454	0.7527
3	0.7566	0.7495
4	0.7431	0.7414
5	0.7548	0.0000
6	0.7563	0.0000
7	0.7521	0.0000
8	0.7457	0.0000
9	0.7508	0.0000

Figure 4.24: client accuracy in async

Table 4.3: Summary of Client Training Accuracy Experiments for Sync and Async Strategies

Aspect	**Asynchronous Strategy (Experiment 1)**	**Synchronous Strategy (Experiment 2)**
Number of Clients	5	10
Initial Accuracy	0.4	0.37
Final Accuracy	0.7	0.74
Accuracy Evolution	Gradual improvement over time	Faster convergence to higher accuracy
Result Variability Among Clients	High variability due to communication/data differences	Uniform performance among clients
Client Accuracy per Round	Line plot with dashed colored lines for each of the 5 clients; all clients show improvement	Line plot with solid colored lines for each of the 10 clients with dots; clients show cohesive improvement
Sync vs Async Accuracy Comparison	Orange line (Async) shows slower increase with X markers	Blue line (Sync) shows higher accuracy with circular markers
Insights	Potential discrepancies due to network latency or data heterogeneity	Consistent updates lead to uniform and stable learning

The results showed that the synchronous strategy (Sync) achieved a higher final accuracy (0.74) compared to the asynchronous strategy (Async), which stopped at 0.70.

In the synchronous mode, the server aggregates updates from all clients before updating the global model. This approach ensures comprehensive data utilization and produces a more balanced and stable model, reflected in the higher final accuracy and consistent client performance, where client accuracies ranged between 0.73 and 0.75, indicating clear stability in collective learning.

In contrast, the asynchronous mode updates the model as soon as any client's update arrives, leading to faster updates but potentially causing fluctuations in performance due to incomplete or heterogeneous data usage. This was evident in the variation of client accuracies, which ranged from 0.68 to 0.72, suggesting differences in local data quality or communication speeds.

Synchronous federated learning provides more stable performance and higher collective accuracy when a reliable and regular communication environment is available. Meanwhile, asynchronous federated learning is preferable in heterogeneous environments or when

faster training is required, despite the possible variance in client performances.

4.6 Interpreting Models Using GradCAM and NormGrad

The aim of this experiment is to understand how the Convolutional Neural Network (CNN) model makes decisions by analyzing the regions it focuses on during prediction.

To achieve this, the GradCAM and NormGrad techniques were applied, which generate heatmaps that highlight the most influential parts of the image during the model's decision-making process.

This experiment seeks to analyze how well the model focuses on meaningful visual features within the images, helping to interpret the model's behavior and assess whether it relies on relevant information when making predictions.

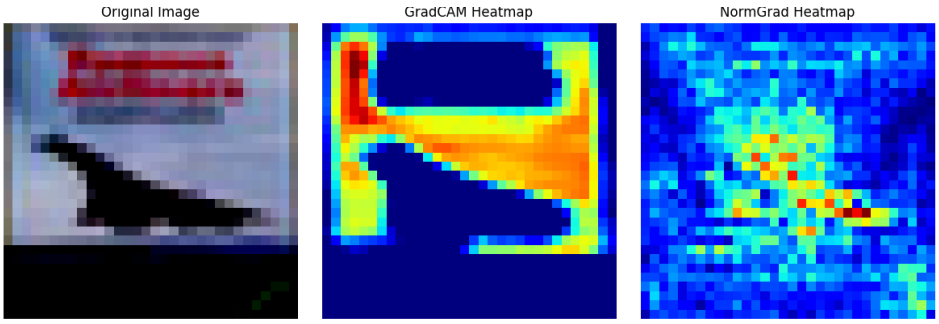

Figure 4.25: Interpretation of image 1

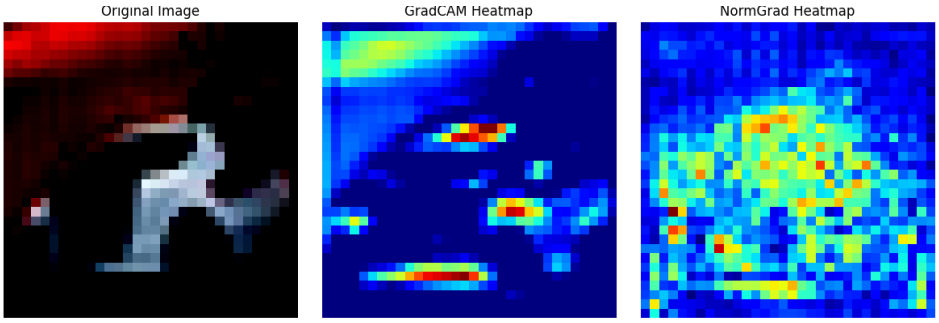

Figure 4.26: Interpretation of image 2

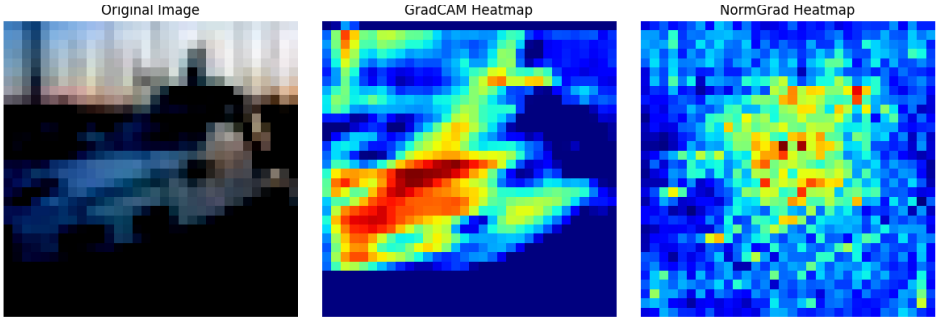

Figure 4.27: Interpretation of image 3

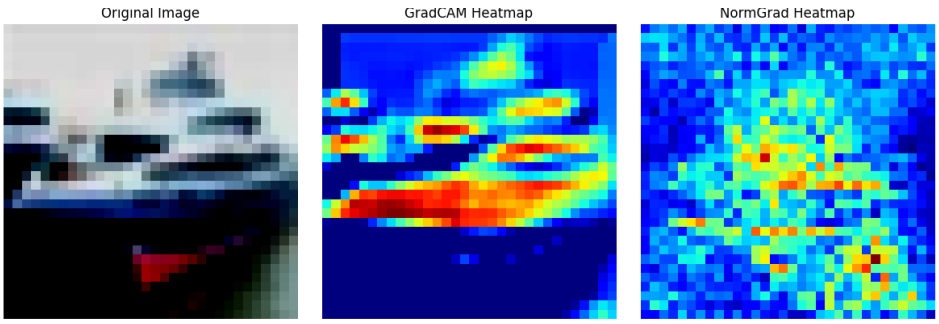

Figure 4.28: Interpretation of image 4

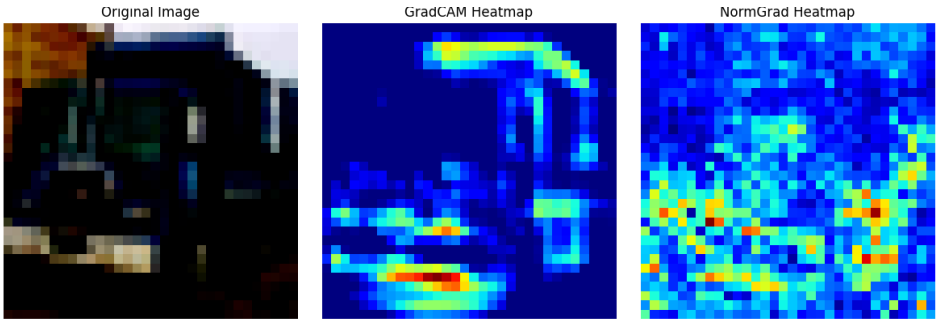

Figure 4.29: Interpretation of image 5

The Results and Visual Analysis

Visual Interpretation	Description
Original Image	The image passed to the model. It contains a low-resolution object with complex background elements.
GradCAM Heatmap	The heatmap highlights the most important regions using warm colors (red/yellow), showing areas where the model focused.
NormGrad Heatmap	The NormGrad heatmap provides a finer, more localized interpretation, emphasizing smaller but more concentrated areas.

Table 4.4: Visual Interpretation of the CNN Model's Attention

- **GradCAM**: The heatmap showed broad areas that roughly align with the shape of the object. The model focused on large regions highlighting the most important parts of the image.
- **NormGrad**: The heatmap provided more detailed focus, indicating that the model was able to pinpoint smaller, more concentrated areas that were deemed important.

These techniques provide valuable insights into how the model makes decisions. **GradCAM** offers a broad understanding of where the model is focusing its attention, while **NormGrad** provides a finer and more precise focus on smaller, more specific features.

These techniques contribute to improving the interpretability of the model, which enhances transparency, especially in privacy-sensitive federated learning environments. In federated learning applications dealing with sensitive data, such as healthcare, the transparency in model decision-making becomes increasingly important. Techniques such as **GradCAM** and **NormGrad** help ensure that the model is interacting with data in a trustworthy and fair manner.

4.7 The Impact of Privacy-Preserving Noise on Model Performance and Interpretability in Federated Learning

To examine how privacy-enhancing techniques affect both model performance and interpretability in a federated learning environment, Gaussian noise was injected into model gradients during local training. This approach simulates differential privacy mechanisms and aims to mitigate the risk of sensitive information leakage during communication.

The experiments were conducted using a Convolutional Neural Network (CNN) model under the standard federated learning setup. Privacy noise levels were varied across four configurations:

- **Noise Level 0.00:** No noise added (baseline).
- **Noise Level 0.01:** Minimal privacy noise.
- **Noise Level 0.05:** Moderate privacy noise.
- **Noise Level 0.10:** High privacy noise.

GradCAM and NormGrad were used to extract visual interpretations from the trained client models according to each privacy setting. The accuracy and quality of interpretation of the results were then evaluated.

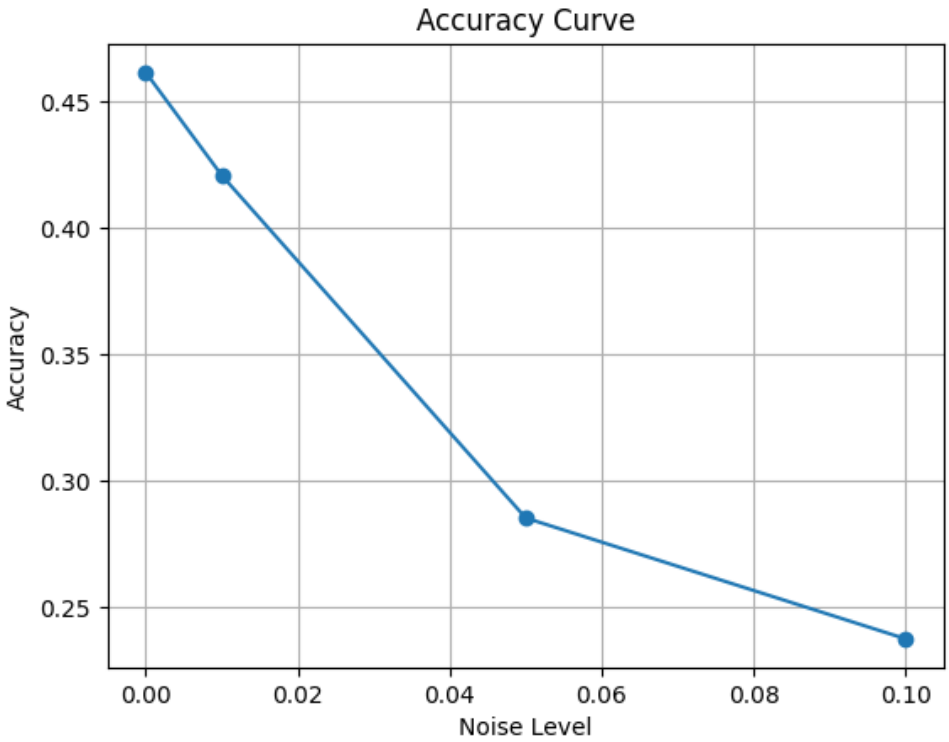

Figure 4.30: Accuracy during Noise Level

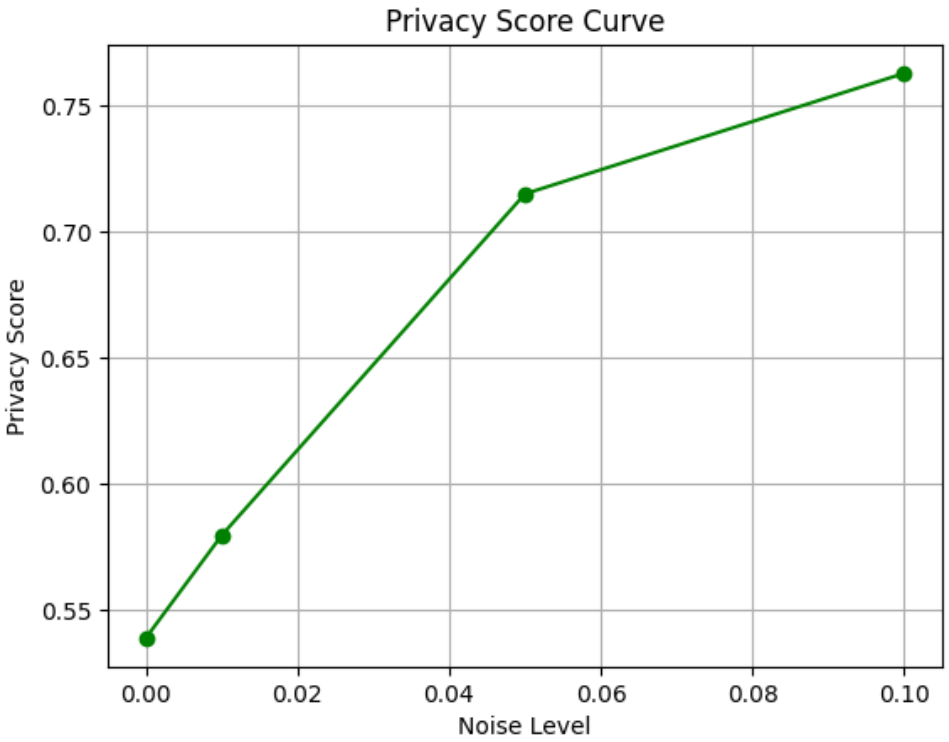

Figure 4.31: Privacy Score during Noise Level

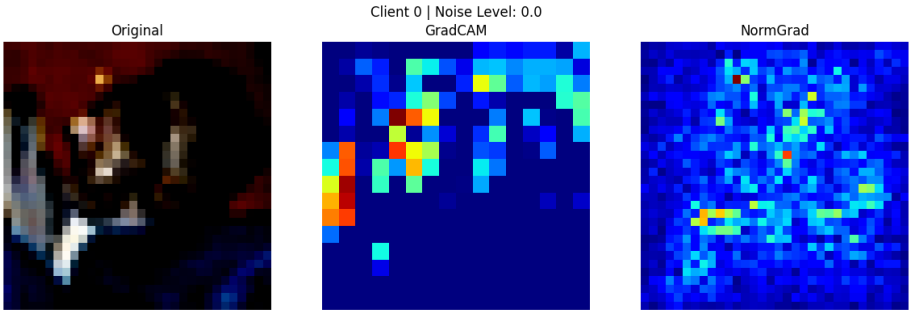

Figure 4.32: Interpretability in 0 Noise Level

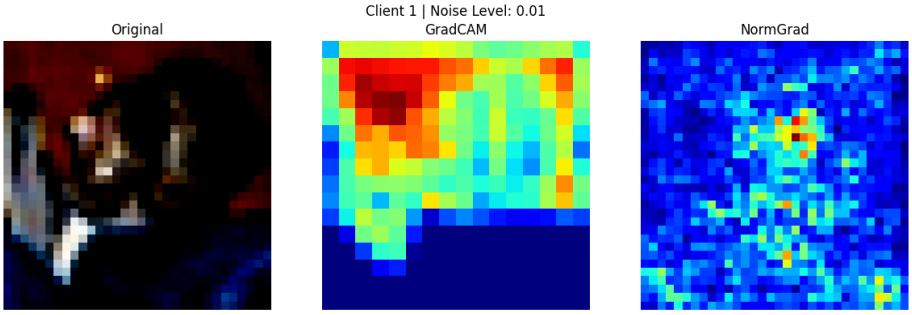

Figure 4.33: Interpretability in 0.1 Noise Level

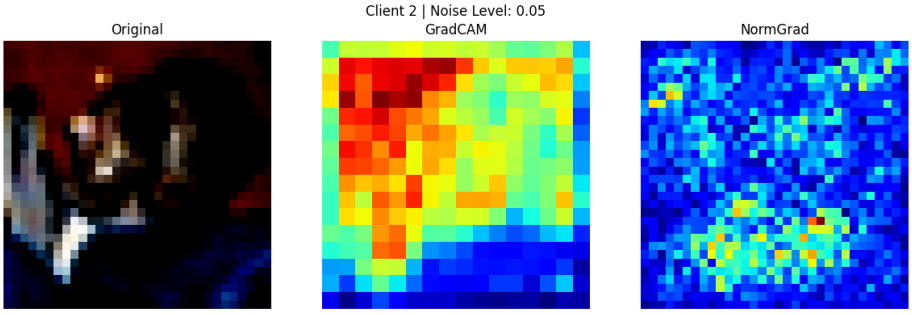

Figure 4.34: Interpretability in 0.5 Noise Level

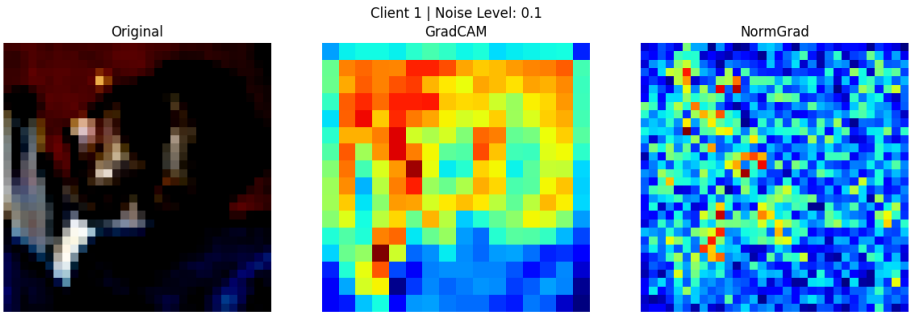

Figure 4.35: Interpretability in 1 Noise Level

The following table summarizes the observed effects of increasing noise levels on both model performance and interpretability:

Table 4.5: Impact of Privacy Noise on Model Accuracy and Visual Explanations

Noise Level	**Accuracy**	**GradCAM Clarity**	**NormGrad Sharpness**	**Privacy Score**
0.00	High	Focused and clear regions	Accurate and stable details	Low
0.01	Slightly reduced	Expansion of important areas	Relatively stable and good detail	Slight increase
0.05	Moderate	Greater attention dispersion	Some change, but less than GradCAM	Moderate
0.10	Low	Clear loss of focus	Limited dispersion	High

The analysis reveals a clear trade-off between privacy and model utility:

- **Performance vs. Privacy:** As the noise level increases, the accuracy of the model consistently declines due to distorted gradient signals, which impede effective learning and generalization.
- **Interpretability Degradation:** Visual explanations generated by GradCAM and NormGrad become less coherent and informative with higher noise levels. This suggests that privacy noise not only disrupts model weights but also corrupts the internal representations critical for explainability.
- **Privacy Score Proxy:** We define a simple privacy proxy metric as Privacy Score = 1 - Accuracy, indicating that higher uncertainty (i.e., lower accuracy) may reflect stronger privacy protection. While simplistic, this aligns qualitatively with differential privacy principles.

The results demonstrate that achieving a balance between privacy and model accuracy in federated learning environments requires careful consideration. Injecting noise into model updates enhances privacy protection but leads to a decrease in predictive accuracy and the quality of visual model interpretations. Therefore, privacy-preserving techniques should be designed with two main objectives in mind: protecting sensitive data and ensuring model reliability.

4.8 Impact of Model Interpretability using GradCAM and NormGrad in Standard and Multi-task Federated Learning

To study the impact of visual interpretation techniques in analyzing the behavior of models trained in a federated learning environment, we applied **GradCAM** and **NormGrad** to extract attention maps from both local and global models after each round. The experiment was conducted using a simple CNN model under two different settings:

- **Standard Federated Learning (Single-task FL)**: All clients participate in training a single classification task.
- **Multi-task Federated Learning (Multi-task FL)**: The model learns from multiple tasks coming from diverse client data.

The experimental settings were kept consistent across both scenarios (same number of clients, rounds, optimization strategy, and model architecture) to ensure a fair comparison.

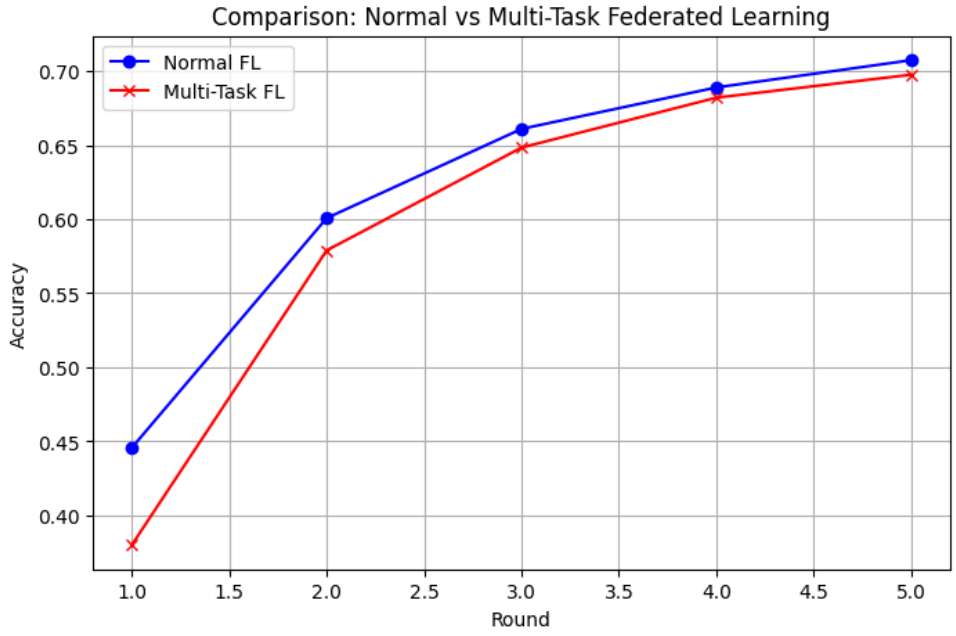

Figure 4.36: accuracy of Standard and Multi-task Federated Learning

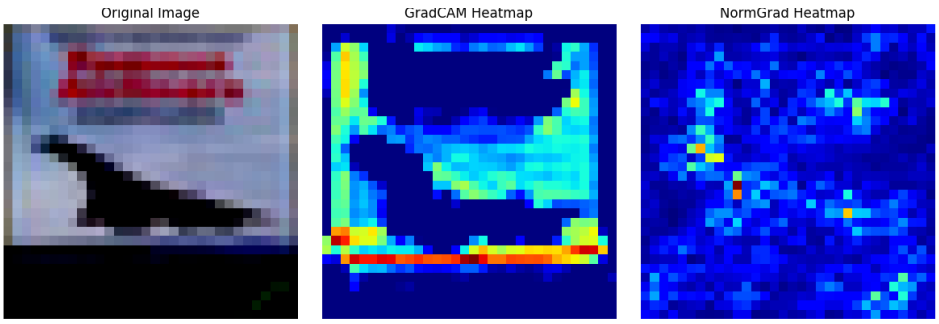

Figure 4.37: Interpretation of Standard Federated Learning

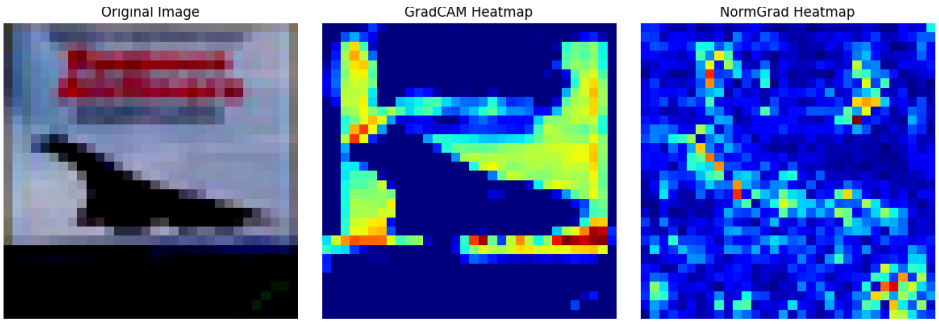

Figure 4.38: Interpretation of task Federated Learning

GradCAM and NormGrad were used to generate heatmaps from the same sample image for both models from a given client. Table 4.8 presents a qualitative comparison of interpretability between the two approaches.

Table 4.6: Qualitative comparison between standard and multi-task federated learning using GradCAM and NormGrad.

Metric	**Standard Federated Learning**	**Multi-task Federated Learning**
GradCAM Clarity	Focuses on object edges (e.g., airplane) and red lines, with moderate attention elsewhere.	Stronger and more precise focus on the main object, with better feature highlighting.
GradCAM Distribution	Attention is localized in few regions, missing some important parts.	Broader attention covers more of the object, indicating deeper understanding.
NormGrad Clarity	Sparse and scattered highlight points, with limited informative detail.	Bright spots are few in density and distribution.
Interpretation	Model relies on coarse features for decision-making without comprehensive understanding.	Model captures more advanced features, showing richer and more interpretable learning.
Accuracy	70.7%	69.7%

The results show that **multi-task federated learning** significantly improves the model's ability to interpret and understand images. This can be explained by the following insights:

- **Multi-tasking Promotes General Representations:** In the multi-task setting, the model is exposed to more diverse tasks and data distributions, encouraging it to learn deeper and more generalizable features, which directly reflect in the attention maps.
- **Focus on Discriminative Features:** In standard FL, the model might rely on surface-level patterns. However, multi-task FL encourages discovery of more robust and task-relevant features, resulting in higher interpretability.
- **Interpretability as a Complement to Accuracy:** Although the accuracy of the multi-task model (69.7%) was slightly lower than the single-task model (70.7%), the multi-task model demonstrated superior visual alignment and semantic clarity in its attention maps. This indicates that interpretability can serve as a complementary metric to accuracy when evaluating model quality and trustworthiness.

Using interpretability tools like GradCAM and NormGrad highlights that **multi-task federated learning models** not only leverage diverse data better but also exhibit **superior visual understanding and decision transparency**. This makes them especially suitable for critical applications such as healthcare, smart surveillance, and sensitive predictive systems.

4.9 Investigating Weight Sharing Effectiveness in Federated Learning Under Different Data Distributions

This experiment aims to evaluate the **effectiveness of weight sharing between clients** in federated learning. Specifically, it investigates whether clients can benefit from **knowledge sharing through global weight aggregation** under different data distribution scenarios.

To achieve this, the performance of **two evaluation methods** is compared:

- **Method 1:** Evaluation is performed using an external test dataset (*Testset*).
- **Method 2:** Evaluation is conducted using training data from other clients (*Other Trainset*).

The experiment is conducted over **three communication rounds** (Rounds 0, 1, and 2), with accuracy measurements recorded for each method under the following three data distributions:

- **IID:** Each client has a balanced dataset containing samples from all classes.
- **NONIID:** Each client's dataset is restricted to only one or two specific classes.
- **SEMIIID:** Clients possess a dominant class in their data, with a small number of samples from other classes.

The goal is to determine how these distributions affect the ability of the clients to collaborate effectively through shared model updates.

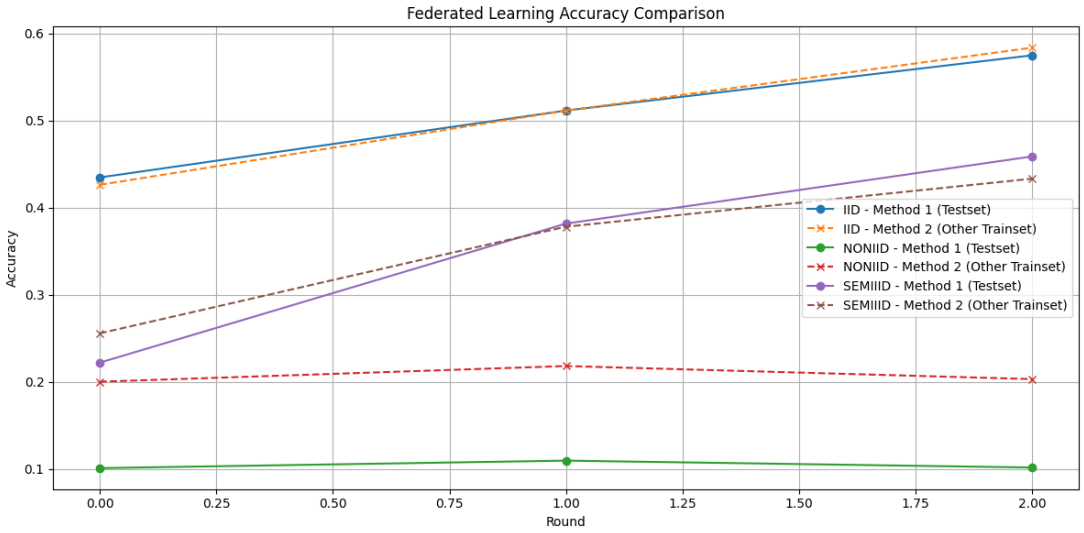

Figure 4.39: accuracy plot

Table 4.7: Accuracy and Weight Sharing Observations Across Data Distributions

Data Distribution	**General Performance**	**Weight Sharing?**	**Key Observations**
IID	Excellent, improving over time	Yes	Accuracy increases steadily. Strong collaborative learning due to shared patterns.
NONIID	Very low, almost static	No	Poor generalization. No effective use of shared weights. Clients learn independently.
SEMIIID	Moderate, gradual improvement	Partial	Some benefits from shared weights, but inconsistency remains compared to IID.

The results show that the performance of federated learning models is highly sensitive to the underlying data distribution:

- In **IID** settings, clients produce balanced updates, leading to effective global model improvement.
- In **NONIID** scenarios, local updates are class-biased and contradict each other during aggregation, causing instability and low accuracy.
- The **SEMIIID** distribution allows for partial alignment between client updates, leading to some improvement, but not as effective as IID.

This confirms that **the quality of data distribution directly affects knowledge sharing and generalization** in federated learning systems.

4.10 Impact of Model Interpretability using GradCAM and NormGrad

To assess the role of interpretability techniques in evaluating and understanding model behavior within federated learning, we applied GradCAM and NormGrad to two distinct CNN architectures: a simple model and a more complex model. Both models were trained under identical federated learning settings, with the same number of clients, data distribution, rounds, and optimization strategy. After each federated round, GradCAM and NormGrad were used to extract heatmaps from each client's local model and from the global model. These heatmaps were analyzed to evaluate decision interpretability and visual attention.

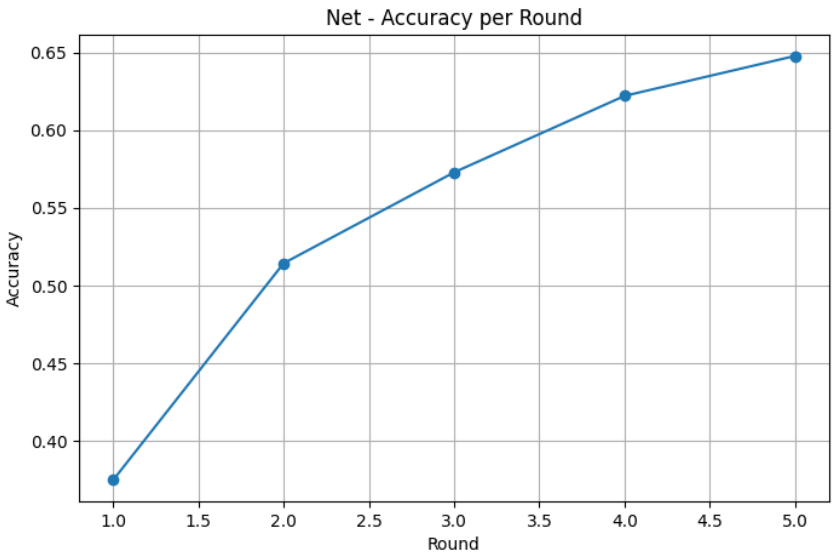

Figure 4.40: accuracy plot Net(1)

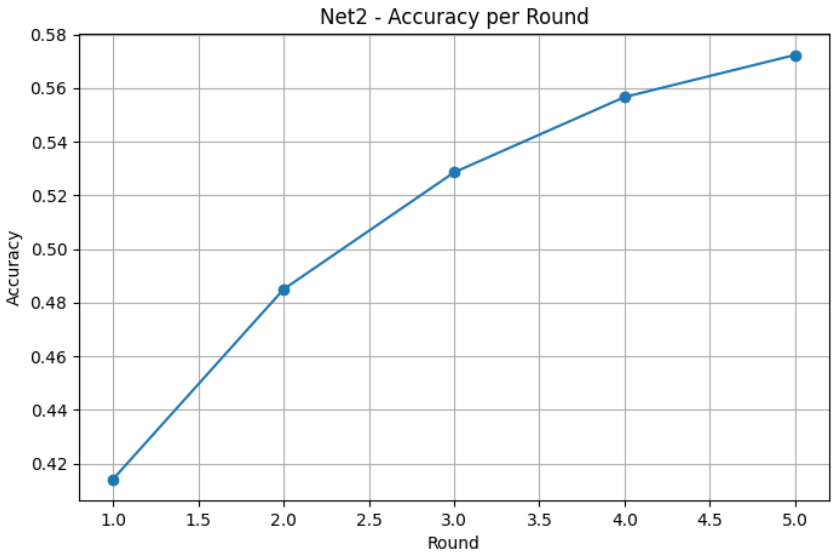

Figure 4.41: accuracy plot Net(2)

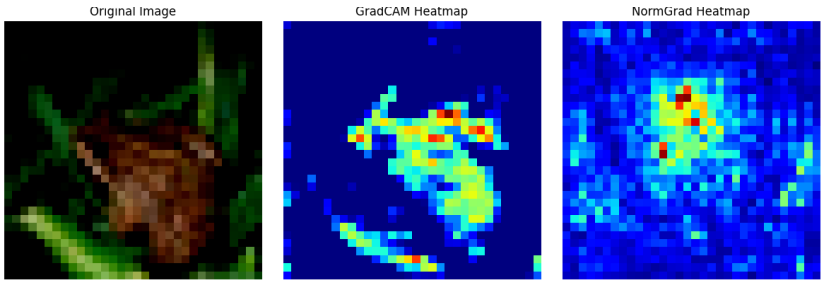

Figure 4.42: Interpretability Net(1)

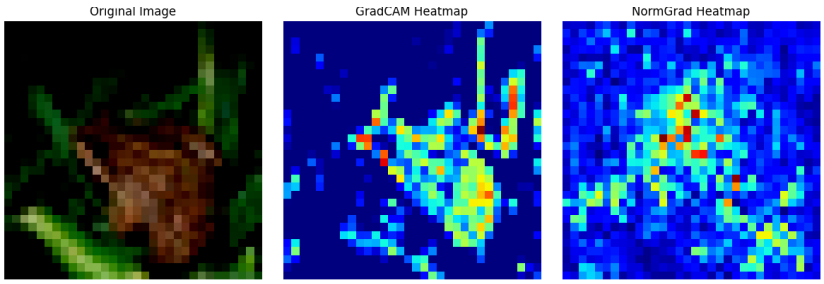

Figure 4.43: Interpretability Net(2)

Model Type	Accuracy	GradCAM Heatmaps	NormGrad Heatmaps
Simple CNN	0.65	Shows light focus areas with color gradients. Highlights important regions, but with limited detail	Displays more detail compared to GradCAM. Higher contrast makes it easier to interpret key areas
Complex CNN	0.57	Sharper and clearer details. Contains vibrant points indicating stronger model focus.	Reveals finer and more extensive details than Simple CNN. Colors are more varied and contrasted, highlighting key components.

Table 4.8: Comparison of simple and complex models using GradCAM and NormGrad.

While the **simple CNN** achieved higher classification accuracy (0.65) compared to the **complex CNN** (0.57), its interpretability was comparatively moderate. The GradCAM and NormGrad heatmaps generated from the simple model provided a general indication of attention regions and offered some level of semantic alignment. However, they lacked the precision and depth observed in the outputs of the complex model.

The **complex CNN**, despite its lower accuracy, demonstrated a stronger ability to localize and highlight meaningful features. GradCAM maps showed well-focused attention on key areas such as object contours, while NormGrad outputs were especially rich in detail, contrast, and spatial coverage.

This comparison reveals that **interpretability and accuracy do not always align**. The simple model, though more accurate numerically, produced only moderately interpretable explanations. In contrast, the complex model, through its expressive internal structure, offered clearer and more informative visual insights, making it potentially more valuable for applications requiring transparency and trust.

4.11 Impact of the Number of Clients on Performance and Interpretability

To investigate the impact of the number of clients on the model's performance and decision interpretability in a federated learning environment, four experiments were conducted using different numbers of clients: 2, 5, 10, and 20. All other settings were kept constant, including the model architecture, data configuration, number of federated rounds, and the optimization strategy. GradCAM and NormGrad algorithms were used to generate interpretability heatmaps, and the performance curve were monitored across the federated rounds.

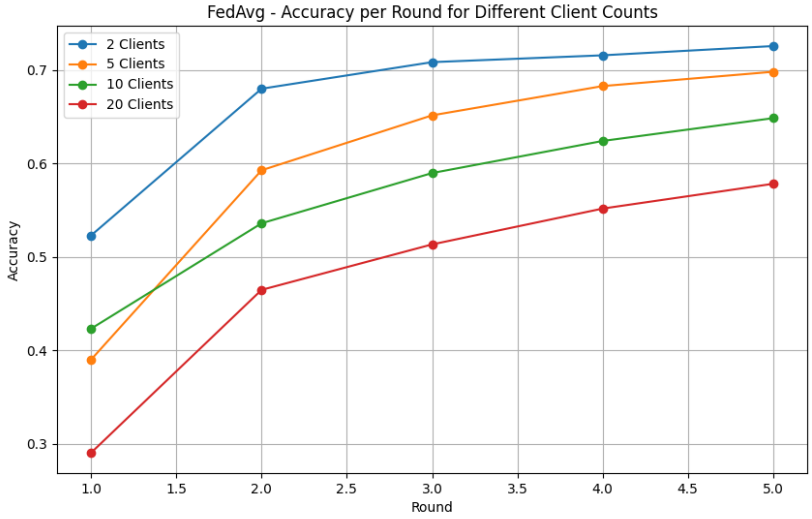

Figure 4.44: plot of accuracy using different numbers of clients: 2, 5, 10, and 20

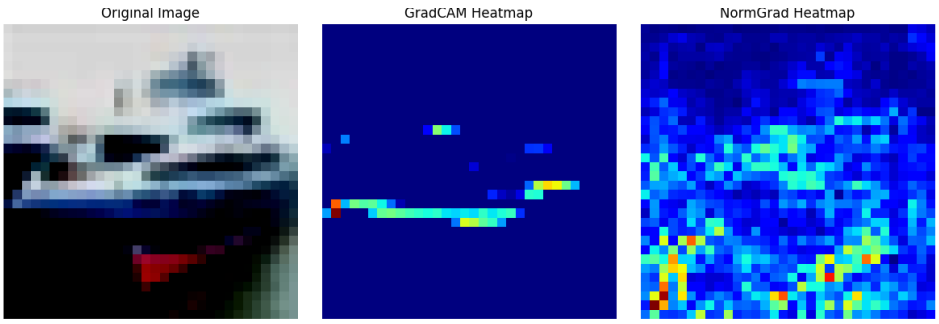

Figure 4.45: Interpretation of a sample image for Client 2.

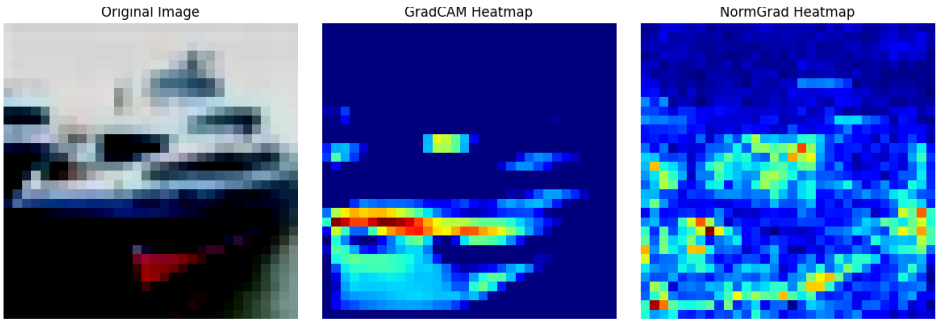

Figure 4.46: Interpretation of a sample image for Client 5.

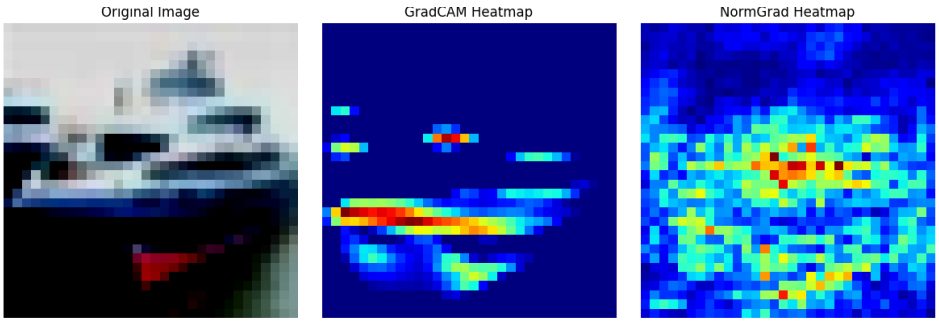

Figure 4.47: Interpretation of a sample image for Client 10.

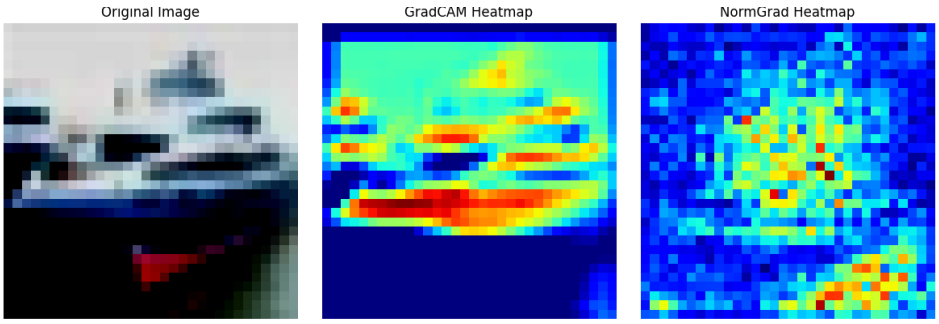

Figure 4.48: Interpretation of a sample image for Client 20.

Number of Clients	Performance (Accuracy)	Heatmap Characteristics
2 Clients	Improvement reaching 0.74	Sparse and weak activations
5 Clients	Improvement reaching 0.70	Emergence of some correlated activations
10 Clients	Improvement reaching 0.65	Clearer and more defined activations
20 Clients	Improvement reaching 0.58	Focused and object-centered activations

Table 4.9: Impact of client number on model performance and heatmap interpretability.

With 2 clients, the heatmaps were scattered and weak. At 5 clients, a slight improvement was observed, with the emergence of some meaningful activation patterns, although less clear than with 10 or 20 clients. With 10 and 20 clients, heatmaps became much more focused and centered on relevant objects.

Although the theoretical premise suggests that increasing the number of clients in a federated learning environment enhances data diversity and improves the model's generalization capability, practical experiments have shown a noticeable decline in model accuracy as the number of clients increases.

This decline is attributed to each client receiving a very small portion of the data when it is distributed among a large number of clients, resulting in weak local learning due to insufficient data. Consequently, this excessive partitioning produces unstable and low-quality updates, which complicate the aggregation process at the central server and negatively impact model convergence.

Despite this quantitative decline in performance, heatmaps generated using interpretability tools such as GradCAM and NormGrad have shown a significant improvement in focus and visual precision, particularly in experiments involving 10 and 20 clients.

This improvement is attributed to the increased diversity in visual patterns that the model is exposed to due to the wider data distribution, even if each client holds only a small share. The presence of slight variations in client data enhances the ability of deep network layers to learn more distinctive and comprehensive feature representations, leading to more focused activations on the target objects within images.

Therefore, this diversity—although scattered—enriches the model representationally and supports improved visual interpretability. This underscores the importance of achieving a carefully considered balance between the number of clients and the amount of local data to ensure integration between quantitative performance and interpretability accuracy.

4.12 Physical Resources and Federated Learning Metrics

The objective of this experiment is to evaluate the performance of the selected federated learning strategy in a real-world setting. The evaluation focuses on accuracy, communication efficiency, resource utilization (RAM and CPU), and total training time. The ultimate goal is to measure the effectiveness and efficiency of the federated system in a distributed environment where multiple clients collaborate without directly sharing their data.

The federated learning experiment was conducted using a model trained on data that was evenly distributed among ten clients. Each client trained locally and then sent its updates to a central server. The Flower framework was utilized to implement the federated training with synchronous strategy and balanced data distribution.

Performance was monitored based on the following metrics:

- **Accuracy:** Measures the ability of the global model to correctly classify data.
- **Error Rate:** Represents the proportion of incorrect predictions made by the model.
- **Communication Efficiency:** Total number of bytes transmitted during the federated training rounds.
- **Message Exchange:** Number of messages exchanged between clients and the server.
- **Resource Utilization (RAM and CPU):** Tracks the consumption of computational resources.
- **Global Training Time:** Total time required to complete all training rounds.

Table 4.10: Final Summary of Federated Learning Performance Metrics

Metric	Value
Average Accuracy	0.7181
Error Rate	0.2819
Average Communication Efficiency (Bytes)	339,018,640
Average Message Exchange	200
Average RAM Usage (%)	37.7 %
Average CPU Usage (%)	9.0 %
Number of Rounds	10
Number of Participants	10
Global Training Time (seconds)	7780.61

- An accuracy of **71.81%** indicates that the federated model successfully learned from distributed data without requiring central aggregation, achieving reasonably good classification performance.

- An error rate of **28.19%** is acceptable given the distributed nature of the data and the possible asynchrony in updates.
- The relatively high communication cost of approximately **339MB** reflects the overhead of transmitting model updates between clients and the central server over multiple rounds, which is expected in federated setups.
- An average of **200 messages** exchanged demonstrates active participation in each round, representing a fully executed federated process.
- **Low RAM (37.7%) and CPU usage (9.0%)** show that the implementation is resource-efficient and does not significantly burden the participating devices.
- A total training time of **7780.61 seconds** (approximately 2.16 hours) suggests that federated learning requires more time than centralized training, but it offers benefits such as data privacy and load distribution.

General Conclusion

This research has deeply addressed a central issue in modern artificial intelligence: how to achieve a delicate balance between privacy and transparency in a federated learning environment where sensitive data is distributed across multiple devices. By combining Federated Learning (FL) and Explainable Artificial Intelligence (XAI), this project aimed to provide an integrated framework that supports intelligent decision-making in Internet of Things (IoT) applications, without compromising data security or neglecting the need for understanding model decisions.

The experiments demonstrated that although FL provides substantial advantages in preserving data privacy, it faces essential challenges in explaining model decisions. However, the integration of visual interpretability tools such as GradCAM and NormGrad enabled the identification of the most influential areas in the decision-making process without breaching privacy boundaries. These tools offered clear and convincing explanations, especially in scenarios involving non-uniform data distributions—representing a significant advancement in this field.

The comparative analysis of FL strategies—**FedAvg**, **FedAdam**, **FedOpt**, and **FedYogi**—under a uniform and synchronous data distribution revealed consistent yet nuanced variations in performance. **FedAvg** and **FedOpt** demonstrated stable and balanced results across clients, with **FedOpt** showing slightly better communication efficiency. **FedAdam** exhibited higher sensitivity to local updates, resulting in greater variability in client accuracy, even in this homogeneous context. **FedYogi** maintained moderate performance, reflecting its design for non-IID scenarios, but did not show significant advantages under balanced conditions.

The real value of this work lies in its practical and tested approach, which unifies privacy, efficiency, and interpretability—making the findings applicable to real-world systems such as smart healthcare, environmental monitoring, and decentralized financial analysis. The framework was built with the constraints of low-resource systems in mind, paving the way for deployment in realistic and complex scenarios.

Although this research constitutes a meaningful step in the right direction, there remains ample room for further development and enhancement. Future work should focus on developing lightweight federated interpretability techniques suitable for resource-constrained IoT environments, investigating the impact of malicious attacks on interpretability, and integrating mechanisms for verifying the authenticity of updates in FL contexts. Moreover, applying the framework to real medical or financial datasets will further validate its robustness and broaden its applicability.

In conclusion, this project represents a foundational step toward building responsible AI models that neither compromise on privacy nor disregard the need for transparency. It is a call to transition from opaque "black-box" AI systems to transparent, fair, and secure intelligence—grounded in trust and awareness—and aligned with the future we aspire to shape.

Bibliography

[1] Konecny, J., McMahan, H. B., Yu, F. X., Richtarik, P., Suresh, A. T., & Bacon, D. (2016). Federated Learning: Strategies for Improving Communication Efficiency. arXiv preprint arXiv:1610.05492.

[2] Sheller, M. J., Reina, G. A., Edwards, B., Martin, J., & Bakas, S. (2020). Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific Reports*, 10(1), 12598.

[3] 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.

[4] Hard, A., Rao, K., Mathews, R., Ramaswamy, S., Beaufays, F., Augenstein, S., ... & McMahan, H. B. (2018). Federated learning for mobile keyboard prediction. arXiv preprint arXiv:1811.03604.

[5] Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. arXiv preprint arXiv:1702.08608.

[6] Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, 30, 4765–4774.

[7] Shokri, G., & Shmatikov, V. (2021). Privacy Risks of Model Explanations. In *IEEE Symposium on Security and Privacy (SP)*, 210–226.

[8] Abadi, M., et al. (2016). Deep Learning with Differential Privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 308–318.

[9] Chamikara, R., Bertok, P., Khalil, I., Liu, D., & Camtepe, S. (2020). Privacy Preserving Machine Learning with Federated Learning and Differential Privacy. *IEEE Transactions on Information Forensics and Security*, 15, 2690–2701.

[10] Ribeiro, M. T., Singh, S., & 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*, 1135–1144.

[11] Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. arXiv preprint arXiv:1409.0473.

[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] IBM. (2023). The State of AI and Privacy: Trends in Federated Learning Adoption. IBM Research Report.

[14] European Commission. (2022). Privacy and Data Security in AI: The Role of Federated Learning. EU Policy Brief.

[15] GDPR Compliance Board. (2021). Data Privacy Regulations and Emerging AI Technologies. *Legal AI Journal*, 18(4), 145–163.

[16] Zhang, X., Wang, T., & Yang, F. (2022). Interpretable AI in Medical Diagnosis: A Federated Learning Perspective. *Journal of Medical AI Research*, 15(2), 87–102.

[17] McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

[18] Fung, C., Yoon, C. J., & Beschastnikh, I. (2018). Mitigating Sybils in federated learning poisoning. arXiv preprint arXiv:1808.04866.

[19] Kairouz, P., McMahan, H. B., Avent, B., et al. (2021). Advances and Open Problems in Federated Learning. *Foundations and Trends in Machine Learning*.

[20] Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., & Shmatikov, V. (2020). How to backdoor federated learning. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*.

[21] Sun, X., Kairouz, P., Suresh, A. T., & McMahan, H. B. (2019). Can you really backdoor federated learning? arXiv preprint arXiv:1911.07963.

[22] Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS)*.

[23] Shokri, R., & Shmatikov, V. (2015). Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS)*.

[24] Beutel, D. J., Topal, T., Mathur, A., Qiu, X., Parcollet, T., Wang, Y., & Lane, N. D. (2022). Flower: A friendly federated learning framework. *arXiv preprint arXiv:2007.14390*.

[25] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[26] Craven, M., & Shavlik, J. (1996). Extracting tree-structured representations of trained networks. In *NeurIPS*.

[27] Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*.

[28] Adel ElZemity, Budi Arief. *Privacy Threats and Countermeasures in Federated Learning for Internet of Things: A Systematic Review*. School of Computing, University of Kent, 2024.

[29] Luis M. Lopez-Ramos et al. *Interplay between Federated Learning and Explainable Artificial Intelligence: A Scoping Review*. VALIDATE Consortium, 2024.

[30] Nguyen Truong, Kai Sun, Siyao Wang, Florian Guitton, YiKe Guo. *Privacy Preservation in Federated Learning: An Insightful Survey from the GDPR Perspective*. Imperial College London, 2021.

[31] Sofia Zahri, Hajar Bennouri, Ahmed M. Abdelmoniem. *An Empirical Study of Efficiency and Privacy of Federated Learning Algorithms*. Queen Mary University of London, TU-Dublin, 2024.

[32] Pedro Miguel Sánchez Sánchez et al. *FederatedTrust: A Solution for Trustworthy Federated Learning*. University of Murcia, University of Zurich, Cyber-Defence Campus, 2023.

[33] Si-ahmed Ayoub, Al-Garadi Mohammed Ali, Boustia Narhimene. *Explainable Machine Learning-Based Security and Privacy Protection Framework for Internet of Medical Things Systems*. Blida 1 University, Emory University, 2024.

[34] Anichur Rahman et al. *Federated Learning-Based AI Approaches in Smart Healthcare: Concepts, Taxonomies, Challenges, and Open Issues*. National Institute of Textile Engineering and Research, King Saud University, Aalto University, 2023.

[35] Anmin Fu, Xianglong Zhang, Naixue Xiong, Yansong Gao, Huaqun Wang. *VFL: A Verifiable Federated Learning with Privacy-Preserving for Big Data in Industrial IoT*. Nanjing University of Science and Technology, Northeastern State University, 2020.

[36] Rémi Gosselin, Loïc Vieu, Faiza Loukil, Alexandre Benoit. *Privacy and Security in Federated Learning: A Survey*. Appl. Sci. 2022, 12, 9901. https://doi.org/10.3390/app12199901.

[37] Tien-Dung Cao, Tram Truong-Huu, Hien Tran, Khanh Tran. *A Federated Learning Framework for Privacy-preserving and Parallel Training*. Tan Tao University, Vietnam; A*STAR, Singapore, 2021.

[38] Fardin Jalil Piran, Zhiling Chen, Mohsen Imani, Farhad Imani. *Privacy-Preserving Federated Learning with Differentially Private Hyperdimensional Computing*. University of Connecticut, University of California Irvine, 2024.

[39] Tien-Dung Cao, Tram Truong-Huu, Hien Tran, Khanh Tran. *A Federated Deep Learning Framework for Privacy Preservation and Communication Efficiency*. Tan Tao University, Singapore Institute of Technology, Vietnam National University, 2022.

[40] Ali Raza. *Secure and Privacy-preserving Federated Learning with Explainable Artificial Intelligence for Smart Healthcare System*. Ph.D. Thesis, University of Lille & University of Kent, 2023.

[41] Alessandro Renda et al. *Federated Learning of Explainable AI Models in 6G Systems: Towards Secure and Automated Vehicle Networking*. Information 2022, 13, 395.

[42] S. S. Pradeep, R. K. Gupta, and A. K. Mishra. *Federated Learning for Privacy-Preserving Healthcare Analytics: Opportunities and Challenges*. Int. J. Healthcare Inf. Syst. Informatics, vol. 21, no. 4, pp. 1–15, 2023. .

[43] José Luis Corcuera Bárcena et al. *Fed-XAI: Federated Learning of Explainable Artificial Intelligence Models*. University of Pisa, 2022.

[44] Tomisin Awosika et al. *Transparency and Privacy: The Role of Explainable AI and Federated Learning in Financial Fraud Detection*. Anglia Ruskin University, 2023.

[45] Ghazaleh Shirvani, Saeid Ghasemshirazi, Mohammad Ali Alipour. *Enhancing IoT Security Against DDoS Attacks through Federated Learning*. Preprint on arXiv, March 2024.

[46] Yanna Jiang et al. *Blockchained Federated Learning for Internet of Things: A Comprehensive Survey*. ACM Comput. Surv., 2024.

[47] Wael Issa et al. *Blockchain-based Federated Learning for Securing Internet of Things: A Comprehensive Survey*. ACM Comput. Surv., 2023.

[48] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). *Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization*. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 618–626).

[49] Rebuffi, S.-A., & Vedaldi, A. (2020). *There and Back Again: Revisiting Backpropagation Saliency Methods*. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 8839–8848).