

POLITECNICO DI TORINO

POLITECNICO DI TORINO

Master Degree course in ICT For Smart Societies

Master Degree Thesis

Design, Implementation, and Experimental Evaluation of a Privacy-Preserving Federated Learning Framework for Medical Imaging Classification.

Supervisors
Prof. Carla Fabiana Chiasserini
Ing. Sergio Micalizzi

Candidate
Riccardo Fida

Academic Year 2025-2026

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my thesis supervisor, **Professor Carla Fabiana Chiasserini**, for her guidance, availability, and valuable support throughout this thesis. Her feedback and advice were essential in shaping this work and bringing it to completion.

I would also like to sincerely thank my co-supervisor, **Sergio Micalizzi**, for his support, suggestions, and contributions throughout the development of this work.

A special thank you goes to my friends and colleagues, who accompanied me throughout these years of study and made this experience truly unforgettable.

Finally, I would like to express my deepest gratitude to my family for their constant support, patience, and encouragement. Their presence and confidence in me have been fundamental throughout my academic journey and in reaching this important milestone.

Abstract

Healthcare is one of the most promising domains for the application of Artificial Intelligence (AI). However, the effective use of healthcare data is often limited by fragmentation across institutions, interoperability issues, and strict privacy requirements. Federated Learning (FL) has emerged as a collaborative learning paradigm that enables multiple organizations to train machine learning models without directly sharing sensitive patient data. Despite its advantages, FL remains vulnerable to privacy leakage and security threats that may compromise both data confidentiality and model integrity.

This thesis investigates privacy-preserving federated learning in healthcare, focusing on the effectiveness of Differential Privacy (DP) and Homomorphic Encryption (HE) as mechanisms for enhancing privacy and security. The study analyzes three representative threats: gradient inversion attacks, which attempt to reconstruct training samples from shared gradients; membership inference attacks, which seek to determine whether a specific sample participated in training; and backdoor attacks, in which malicious clients manipulate local data to alter the behavior of the global model.

An experimental framework is developed to evaluate the impact of these attacks and defenses under different federated learning configurations. The evaluation considers both security and performance aspects, including model utility, convergence behavior, computational overhead, communication costs, and resource consumption. The results highlight the trade-offs between privacy protection and learning performance, showing how different defense mechanisms influence model accuracy, efficiency, and robustness.

The findings provide insights into the practical challenges of deploying federated learning in healthcare and assess the feasibility of adopting the proposed privacy-preserving framework in real-world clinical environments while maintaining acceptable performance, scalability, and security guarantees.

Contents

1 Introduction 7
- 1.1 Background and Motivation . . . 7
- 1.2 Problem Statement . . . 8
- 1.3 Research Objectives and Questions . . . 8
- 1.4 Scope and Limitations . . . 9
- 1.5 Main Contributions . . . 9
- 1.6 Thesis Structure . . . 9

2 State of the art 11
- 2.1 Healthcare as a domain for Artificial Intelligence . . . 11
  - 2.1.1 Clinical Data Sources and Healthcare Data Characteristics . . . 12
  - 2.1.2 Sensitive Data, Fragmentation and Interoperability Issues . . . 12
  - 2.1.3 Governance, Trust and Institutional Constraints . . . 13
  - 2.1.4 High-impact Clinical Applications of AI . . . 14
- 2.2 Artificial Intelligence and Machine Learning . . . 15
  - 2.2.1 Basic concepts of artificial intelligence . . . 15
  - 2.2.2 Machine learning and supervised learning . . . 15
  - 2.2.3 Deep learning and data requirements . . . 15
  - 2.2.4 Centralized learning paradigm . . . 16
  - 2.2.5 Limitations of centralized ML in sensitive and distributed environments . . . 16
- 2.3 Federated Learning: general concepts . . . 17
  - 2.3.1 Motivation for federated learning . . . 17
  - 2.3.2 FL architecture and workflow . . . 18
  - 2.3.3 Roles of server and clients . . . 18
  - 2.3.4 Model aggregation and communication rounds . . . 19
  - 2.3.5 Types of Federated Learning . . . 19
  - 2.3.6 Existing frameworks for federated learning . . . 21
- 2.4 Federated Learning in healthcare . . . 22
  - 2.4.1 Why FL is suitable in healthcare . . . 22
  - 2.4.2 FL across hospitals and clinical institutions . . . 23
  - 2.4.3 Healthcare applications: diagnosis, imaging, prediction and monitoring . . . 23

2.4.4 Challenges in healthcare FL: heterogeneity, bias, communication, validation . . . 24
2.4.5 Evaluation, deployment, and clinical integration issues . . . 25
2.5 Legal and regulatory aspects for AI and Federated Learning in healthcare 25
2.5.1 Sensitive Health Data and Legal Obligations . . . 25
2.5.2 GDPR and Data Protection Requirements . . . 26
2.5.3 HIPAA and Healthcare Data Governance . . . 27
2.5.4 AI-specific Risks in Healthcare . . . 27
2.5.5 The EU AI Act and High-risk AI Systems . . . 28
2.5.6 Transparency, Accountability and Human Oversight . . . 29
2.6 Privacy and Security Challenges in Federated Learning . . . 30
2.6.1 Cybersecurity perspective and Threat Model . . . 30
2.6.2 Privacy Attacks . . . 32
2.6.3 Integrity Attacks . . . 34
2.6.4 Privacy-Preserving Mechanisms . . . 35
2.6.5 Robust Aggregation Strategies . . . 38

3 Methodology and Experimental Design 41
3.1 Research Methodology . . . 41
3.2 System Architecture . . . 42
3.2.1 Overall Architecture . . . 43
3.2.2 Flower-Based Federated Learning Framework . . . 44
3.2.3 Communication Workflow . . . 44
3.3 Dataset Description . . . 44
3.3.1 Dataset Selection . . . 45
3.3.2 Data Preprocessing . . . 45
3.3.3 Data Partitioning . . . 46
3.4 Model Architecture . . . 47
3.4.1 Neural Network Architecture . . . 47
3.4.2 Training Configuration . . . 48
3.5 Federated Learning Configuration . . . 50
3.5.1 Flower Framework . . . 50
3.5.2 Aggregation Strategies . . . 51
3.5.3 Federated Training Procedure . . . 51
3.6 Proposed Privacy-Preserving Federated Learning Framework . . . 53
3.6.1 Design Objectives . . . 54
3.6.2 Differential Privacy Integration . . . 54
3.6.3 Homomorphic Encryption Integration . . . 55
3.6.4 Secure Training Workflow . . . 56
3.7 Security Threat Model and Attack Implementation . . . 57
3.7.1 Threat Model . . . 58
3.7.2 Membership Inference Attack . . . 58
3.7.3 Frequency-Injection Backdoor Attack . . . 60
3.7.4 Gradient Inversion Attack . . . 62
3.8 Security Evaluation . . . 63

3.8.1 Evaluation Objectives . . . 63
3.8.2 Membership Inference Evaluation . . . 64
3.8.3 Backdoor Evaluation . . . 64
3.8.4 Gradient Inversion Evaluation . . . 64
3.8.5 Security Metrics . . . 64
3.9 Experimental Design . . . 66
3.9.1 Experimental Objectives . . . 66
3.9.2 Baseline Comparison . . . 66
3.9.3 Aggregation Strategy Evaluation . . . 66
3.9.4 Differential Privacy Evaluation . . . 67
3.9.5 Homomorphic Encryption Evaluation . . . 67

4 Results and Discussion 69
4.1 Experimental Overview . . . 69
4.2 Centralized vs Federated Learning . . . 71
4.2.1 Performance Comparison . . . 71
4.2.2 Computational Analysis . . . 72
4.2.3 Communication Analysis . . . 73
4.2.4 Discussion . . . 74
4.3 Impact of Data Distribution and Aggregation . . . 75
4.3.1 IID Distribution . . . 75
4.3.2 Moderate Non-IID ($\alpha = 0.5$) . . . 76
4.3.3 Severe Non-IID ($\alpha = 0.1$) . . . 76
4.3.4 FedAvg vs CKADA . . . 77
4.4 Impact of Differential Privacy . . . 77
4.4.1 Utility Analysis . . . 77
4.4.2 Computational Overhead . . . 78
4.4.3 Discussion . . . 79
4.5 Impact of Homomorphic Encryption . . . 79
4.5.1 Model Performance . . . 80
4.5.2 Communication Overhead . . . 80
4.5.3 Computational Considerations . . . 82
4.5.4 Discussion . . . 82
4.6 Security Evaluation . . . 83
4.6.1 Quantile Membership Inference Attack . . . 83
4.6.2 Backdoor Attack . . . 86
4.6.3 Gradient Inversion Attack . . . 87
4.7 General Discussion . . . 89
4.7.1 Privacy-Utility Trade-off . . . 89
4.7.2 Security-Performance Trade-off . . . 90
4.7.3 Practical Implications for Healthcare . . . 91

5 Conclusion 93
5.1 Summary of Findings 93
5.2 Research Contributions 93
5.3 Limitations 94
5.4 Future Research Directions 94

A Additional Experimental Results 97

Bibliography 99

Chapter 1

Introduction

1.1 Background and Motivation

Artificial Intelligence (AI) has become an increasingly important technology in modern healthcare, supporting clinicians in tasks such as medical image analysis, disease diagnosis, patient monitoring, and treatment planning. Recent advances in deep learning have enabled highly accurate predictive models capable of assisting healthcare professionals in making faster and more informed clinical decisions. However, the successful development of these models depends on the availability of large, diverse, and high-quality datasets.

In practice, healthcare data are rarely stored in a single location. Hospitals, laboratories, and research institutions typically maintain independent medical databases that cannot easily be shared because of legal, ethical, and privacy constraints. Regulations such as the General Data Protection Regulation (GDPR) impose strict requirements on the processing and exchange of sensitive patient information, making centralized data collection increasingly difficult.

Federated Learning (FL) has emerged as a promising solution to this challenge by enabling multiple institutions to collaboratively train a shared machine learning model without exchanging their local datasets. Instead of transmitting raw medical records, each participant performs model training locally and shares only model updates with a central server responsible for constructing the global model.

Although Federated Learning significantly improves data privacy compared with centralized learning, it does not completely eliminate security risks. Recent studies have demonstrated that exchanged model updates may still leak sensitive information or be manipulated by malicious participants, motivating the development of additional privacy-preserving mechanisms capable of protecting both the learning process and the exchanged information.

This thesis investigates the design and experimental evaluation of a privacy-preserving federated learning framework for healthcare that combines Differential Privacy, Homomorphic Encryption, and robust aggregation techniques to improve the confidentiality and robustness of collaborative model training while maintaining acceptable predictive performance.

1.2 Problem Statement

The problem addressed in this study is illustrated in Figure 1.1.

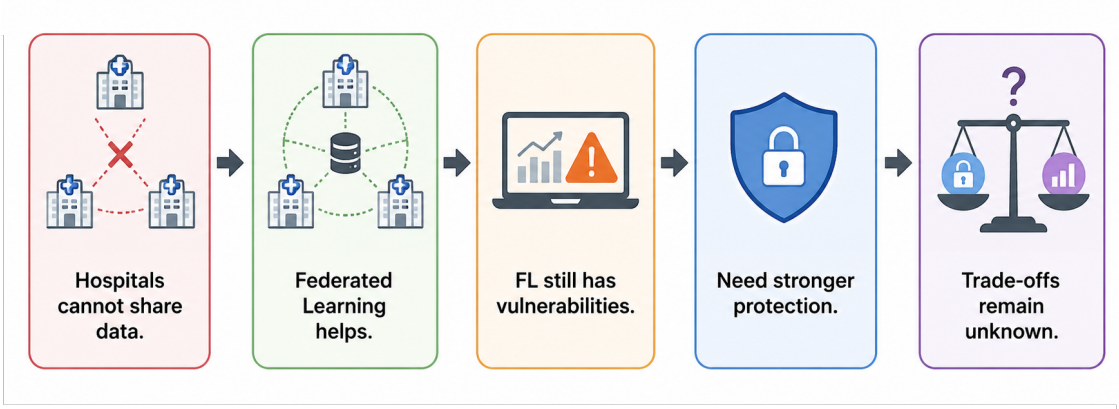


Figure 1.1: Motivation and problem statement for privacy-preserving Federated Learning in healthcare.

1.3 Research Objectives and Questions

The primary objective of this thesis is to design, implement, and experimentally evaluate a privacy-preserving Federated Learning framework for healthcare applications. The proposed framework incorporates Differential Privacy and selective Homomorphic Encryption to strengthen the protection of sensitive medical data while maintaining the predictive performance of collaborative learning. Beyond the implementation of the framework, this work investigates the impact of these mechanisms on model utility, computational efficiency, communication overhead, and resistance to representative privacy and security attacks. Particular attention is devoted to analysing the trade-offs between privacy protection, learning performance, and system efficiency in realistic federated learning scenarios.

To address these objectives, the thesis investigates several research questions. The first research question (RQ1) examines how Federated Learning compares with centralized learning in terms of predictive performance and computational efficiency for healthcare image classification. The second research question (RQ2) investigates the impact of Differential Privacy on model utility and resistance to privacy attacks. The third research question (RQ3) evaluates the communication and computational overhead introduced by selective Homomorphic Encryption.

The fourth research question (RQ4) analyses the vulnerability of the federated learning framework to representative privacy and security attacks, including Membership Inference, Gradient Inversion, and Backdoor attacks. Finally, the fifth research question (RQ5) considers the broader trade-off between privacy protection, security, predictive

performance, and system efficiency when the proposed privacy-preserving mechanisms are applied within the federated learning framework.

1.4 Scope and Limitations

The scope of this thesis is limited to image classification in healthcare using the BloodMNIST dataset [54]. The proposed framework is evaluated in a cross-silo federated learning environment consisting of a single coordinating server and a limited number of participating healthcare institutions. The study focuses on Differential Privacy, Homomorphic Encryption, and robust aggregation as privacy-preserving mechanisms, while considering Membership Inference, Gradient Inversion, and Backdoor attacks as representative adversarial scenarios.

Although the proposed framework provides a comprehensive evaluation of privacy, security, and performance, several aspects remain outside the scope of this work. These include large-scale deployments involving hundreds of clients, fully decentralized federated learning architectures, secure multi-party computation, and real-world clinical deployment. Furthermore, the experimental evaluation is performed using a single medical imaging dataset, and therefore the obtained results may not directly generalize to other healthcare modalities or application domains.

1.5 Main Contributions

This thesis contributes to the development and evaluation of privacy-preserving Federated Learning for healthcare through the design and implementation of a modular framework based on the Flower platform. The proposed system integrates Differential Privacy, Homomorphic Encryption, and hybrid aggregation based on CKADA and FedAvg within a unified training pipeline.

The work also provides an experimental comparison between centralized and federated learning under both IID and non-IID data distributions, allowing the impact of data heterogeneity and distributed training to be assessed in terms of predictive performance and computational efficiency.

A further contribution is the comprehensive security evaluation of the framework against representative threats, including Membership Inference, Gradient Inversion, and Frequency-Injection Backdoor attacks. These experiments provide insight into the effectiveness and limitations of the adopted privacy and aggregation mechanisms under different attack scenarios.

Overall, the thesis offers an integrated analysis of the trade-offs between privacy protection, security, predictive performance, communication cost, and computational overhead, with particular attention to the practical requirements of healthcare federated learning environments.

1.6 Thesis Structure

The remainder of this thesis is organized as follows.

Chapter 2 reviews the theoretical background and the state of the art related to artificial intelligence in healthcare, federated learning, privacy-preserving machine learning, and the main privacy and security threats affecting collaborative learning systems. It also presents the principal defense mechanisms adopted in the literature, including Differential Privacy and Homomorphic Encryption, together with the attacks considered in this work.

Chapter 3 describes the proposed federated learning framework and the experimental methodology. It introduces the system architecture, the BloodMNIST dataset, the data partitioning strategies, the ResNet-18 model architecture, the federated learning configuration, and the implementation of Differential Privacy, Homomorphic Encryption, and hybrid aggregation. The chapter also presents the experimental design, the evaluation metrics, and the implementation of the security attacks considered throughout the study.

Chapter 4 presents and discusses the experimental results. The analysis begins with a comparison between centralized and federated learning, followed by an evaluation of the impact of different data distributions and aggregation strategies. The effects of Differential Privacy and Homomorphic Encryption on model performance, training efficiency, and communication overhead are then investigated. Finally, the security of the proposed framework is evaluated through Membership Inference, Gradient Inversion, and Backdoor attacks, highlighting the trade-offs between predictive performance, privacy, security, and system efficiency.

Chapter 5 concludes the thesis by summarizing the main findings, discussing the limitations of the proposed framework, and outlining possible directions for future research on privacy-preserving federated learning for healthcare applications.

Chapter 2

State of the art

2.1 Healthcare as a domain for Artificial Intelligence

Healthcare has emerged as one of the most promising domains for the application of Artificial Intelligence (AI). The increasing digitalization of healthcare systems, together with the growing availability of clinical data, has created favorable conditions for developing intelligent systems that support medical diagnosis, treatment planning, patient monitoring, and clinical decision-making. Modern healthcare generates large volumes of heterogeneous data, including electronic health records, medical images, laboratory results, genomic information, and data collected through wearable devices. These rich and diverse datasets provide an ideal foundation for machine learning and deep learning techniques [15].

Recent advances demonstrate that AI can significantly improve healthcare delivery by increasing diagnostic accuracy, optimizing clinical workflows, and enabling more personalized treatments. AI models have achieved promising results in several medical specialties, including radiology, oncology, cardiology, drug discovery, and patient risk prediction [15]. By automatically identifying complex patterns within medical data, machine learning algorithms can support earlier disease detection and more accurate prognostic assessments than traditional analytical approaches.

The growing complexity of modern medicine further reinforces the importance of AI. Healthcare professionals must continuously process large amounts of information while making timely and accurate clinical decisions. Rather than replacing clinicians, AI systems are intended to augment clinical expertise by rapidly analyzing heterogeneous datasets and generating evidence-based insights. This capability supports the transition from generalized treatments toward personalized and precision medicine, enabling earlier diagnosis, improved risk prediction, and more effective preventive care. In addition, AI can automate administrative tasks, optimize hospital operations, facilitate remote patient monitoring, and ultimately contribute to more efficient and accessible healthcare services [18].

Despite these significant opportunities, healthcare also presents substantial challenges for AI adoption. Medical information is highly sensitive and subject to strict ethical, legal, and regulatory requirements. Furthermore, healthcare environments are characterized by

fragmented information systems, heterogeneous data formats, and limited interoperability between institutions. Consequently, deploying AI in clinical practice requires not only accurate predictive models but also robust mechanisms that preserve privacy, ensure regulatory compliance, and establish trust among patients, healthcare professionals, and institutions. These challenges directly motivate the need for collaborative and privacy-preserving machine learning approaches, which are discussed throughout this thesis.

2.1.1 Clinical Data Sources and Healthcare Data Characteristics

Healthcare systems generate vast and continuously growing amounts of data originating from multiple heterogeneous sources, including electronic health records (EHRs), hospital information systems, medical images such as CT, MRI, and X-ray scans, laboratory results, genomic information, physiological monitoring devices, and wearable sensors [4]. In addition, healthcare organizations collect physicians' clinical notes, administrative and financial records, patient-generated health data, and information derived from pharmaceutical and scientific research [4]. The widespread adoption of Electronic Medical Records (EMRs) has further accelerated the digitalization of healthcare, enabling the continuous acquisition and storage of patient information across clinical environments [17].

The value of these data lies not only in their volume but also in their diversity. Healthcare data comprise both structured and unstructured information collected from multiple independent systems, often requiring real-time processing and integration. Clinical information may coexist in digital repositories and legacy paper-based records, while differences in data formats and acquisition methods further increase the complexity of analysis [4]. Consequently, conventional data management techniques are often insufficient for extracting meaningful knowledge from these heterogeneous datasets.

Unlike many other application domains, healthcare data are inherently temporal. Patient records consist of sequences of diagnoses, treatments, medications, laboratory tests, and clinical procedures, where the chronological relationships between events play a fundamental role in clinical decision-making [17]. Effectively analyzing these longitudinal patient trajectories therefore requires machine learning techniques capable of capturing temporal dependencies and sequential patterns.

The richness and diversity of healthcare data make the domain particularly well suited to artificial intelligence and advanced analytics. Machine learning, data mining, and predictive analytics can identify clinically relevant patterns, support disease diagnosis, improve risk prediction, and enable personalized medicine [4] [17]. However, despite this considerable potential, the effective use of healthcare data remains constrained by fragmentation across institutions, limited interoperability, and the organizational and technical challenges associated with integrating heterogeneous clinical information. These issues represent one of the principal barriers to the development of robust and generalizable AI models.

2.1.2 Sensitive Data, Fragmentation and Interoperability Issues

Despite generating vast amounts of valuable information, healthcare systems often struggle to exploit these data effectively because they remain fragmented across independent

institutions [35]. Hospitals, clinics, laboratories, pharmacies, insurance providers, and wearable devices typically maintain separate information systems that rarely communicate with one another [4] [17]. As a result, patient medical histories are frequently dispersed across multiple organizations, creating incomplete clinical records and limiting the continuity of care. This phenomenon has also been described as the creation of "shadow health records", where important health information exists outside the primary healthcare system [35].

The diversity of healthcare information systems further complicates data sharing. Clinical data are often stored using different formats, terminologies, and coding standards, making their integration slow, costly, and prone to errors [4] [35]. The COVID-19 pandemic clearly exposed these limitations, as inconsistent definitions and reporting standards hindered the rapid aggregation and exchange of epidemiological data across healthcare organizations [35]. Although initiatives such as the Cures Act Final Rule and the Fast Healthcare Interoperability Resources (FHIR) standard have promoted more interoperable data exchange through standardized application programming interfaces (APIs), their adoption remains heterogeneous across healthcare institutions [35].

Consequently, healthcare organizations continue to operate as isolated data silos, preventing the creation of comprehensive datasets required to train robust and generalizable AI models. While technical interoperability standards have improved data exchange, the effective sharing of healthcare information remains constrained by organizational, legal, and governance challenges, which are discussed in the following subsection.

2.1.3 Governance, Trust and Institutional Constraints

Achieving effective healthcare interoperability requires more than technological compatibility. According to Weber [48], interoperability encompasses four complementary dimensions: technological interoperability, which enables data exchange between systems; data interoperability, which ensures that exchanged information can be correctly interpreted; human interoperability, which facilitates communication among stakeholders; and institutional interoperability, which supports collaboration across independent organizations. While considerable progress has been made in the technological dimension, institutional and regulatory barriers remain among the greatest obstacles to large-scale healthcare data sharing.

Healthcare organizations operate under complex legal and ethical frameworks that regulate the collection, storage, and exchange of sensitive patient information. Privacy legislation varies considerably across countries and jurisdictions and has often failed to keep pace with the rapid evolution of health information technologies [35]. Consequently, organizations face significant legal uncertainty when exchanging data, particularly across institutional or national boundaries, where compliance with one regulatory framework may conflict with the requirements of another.

Beyond legal compliance, effective data sharing also depends on governance and trust. Uncertainties regarding data ownership, stewardship, access rights, and accountability often discourage organizations from sharing clinical information [35]. At the same time, public trust remains fragile, as patients and healthcare providers expect sensitive medical data to be used responsibly and exclusively for legitimate clinical or research purposes.

Any misuse of shared data can rapidly undermine confidence in collaborative healthcare initiatives.

These challenges are closely related to the concept of *legal interoperability*, defined as the ability of legal frameworks to operate coherently across different jurisdictions [48]. Although approaches such as harmonization and mutual recognition have been proposed to facilitate cross-border collaboration, current healthcare systems still exhibit limited legal interoperability. As a result, organizations remain reluctant to exchange sensitive clinical information, restricting access to the large and diverse datasets required for developing robust and generalizable AI models.

The combination of fragmented governance, limited institutional trust, and regulatory constraints therefore represents one of the principal barriers to collaborative artificial intelligence in healthcare. Addressing these limitations while preserving patient privacy is a key motivation for privacy-preserving distributed learning approaches such as Federated Learning, which are introduced in the following section.

2.1.4 High-impact Clinical Applications of AI

Artificial intelligence has achieved remarkable success across a wide range of clinical applications, particularly in medical image analysis and computer-aided diagnosis. Deep learning models have demonstrated excellent performance in the diagnosis and staging of neurological diseases, including Alzheimer's disease, where convolutional neural networks (CNNs), recurrent neural networks (RNNs), and generative adversarial networks (GANs) have been applied to MRI and PET images for disease classification, segmentation, and prognosis [12]. Several studies report diagnostic accuracies exceeding 95% when distinguishing cognitively healthy individuals from patients with mild cognitive impairment and Alzheimer's disease, especially when combining structural MRI with PET imaging biomarkers.

Similarly, AI has shown considerable potential in the diagnosis of Attention Deficit Hyperactivity Disorder (ADHD). Deep learning approaches based on electroencephalography (EEG) and functional magnetic resonance imaging (fMRI) have achieved high diagnostic accuracy, with CNN-based models reporting classification performance of up to 98.48% in pediatric populations [46]. Machine learning techniques have also demonstrated the ability to distinguish ADHD from other neurodevelopmental disorders, such as autism spectrum disorder, thereby supporting more accurate clinical assessments.

Despite these promising results, the widespread adoption of AI in clinical practice remains constrained by several methodological limitations. Many existing studies rely on relatively small, single-center datasets, limiting the generalizability of the developed models to broader patient populations [12] [46]. Although multimodal approaches combining medical images, clinical information, and biomarkers generally improve predictive performance, they also increase the demand for large and diverse datasets collected from multiple healthcare institutions.

Consequently, the principal challenge is no longer the development of increasingly sophisticated deep learning models, but rather providing access to sufficiently large, representative, and heterogeneous clinical datasets while preserving patient privacy and complying with legal and ethical regulations. Addressing this challenge is essential for

developing reliable and generalizable AI systems and constitutes one of the primary motivations for Federated Learning, which enables collaborative model training without requiring healthcare institutions to exchange their raw patient data.

2.2 Artificial Intelligence and Machine Learning

2.2.1 Basic concepts of artificial intelligence

Artificial intelligence (AI) is defined as "an advancement of computing technology that imitates human intelligence by training a program to function humanely." [12]. In practical terms, AI refers to computational systems designed to perform tasks that typically require human-like intelligence, such as pattern recognition, decision-making, prediction, and data interpretation [46]. In healthcare settings, AI has emerged as a powerful tool for enhancing image interpretation, disease classification, and longitudinal prediction of patient outcomes [12]. The fundamental objective of AI in clinical applications is to process large, complex datasets efficiently, identify hidden patterns that human experts may overlook, and generate predictions that support evidence-based clinical decision-making while systematically reducing human bias inherent in traditional diagnostic workflows [46] [10].

2.2.2 Machine learning and supervised learning

Machine learning (ML) is a subset of AI in which systems learn from data by iteratively adjusting internal parameters based on experience rather than following explicitly programmed rules. ML is based on the foundational idea that systems can learn from data, identify patterns, and make decisions with minimal human intervention [10] Within ML, two primary learning paradigms dominate: supervised learning, in which systems receive labeled training data (inputs paired with correct outputs) and learn to predict unknown inputs; and unsupervised learning, where data remains unlabeled and algorithms autonomously discover inherent patterns and clusters [12].

Common supervised learning approaches include support vector machines (SVMs), random forests (RF), logistic regression, and decision trees—all widely applied in healthcare classification tasks. SVMs solve complex classification problems by identifying optimal hyperplanes that separate classes in high-dimensional feature spaces, achieving high accuracy in disease classification when combined with careful feature selection. Random forests demonstrate robust performance across diverse feature sets and clinical modalities [12]. These traditional ML methods remain valuable when training data is limited and interpretability is prioritized over predictive power, though they typically require manual feature engineering by domain experts [10].

2.2.3 Deep learning and data requirements

Deep learning (DL) represents a fundamental advancement beyond traditional ML through its use of artificial neural networks (ANNs)—multilayered networks of interconnected computational nodes that mimic hierarchical processing in biological brains [12]. A deep neural network consists of an input layer, several hidden layers, and an output layer, where

each layer contains multiple units called neurons [10]. Each neuron receives multiple inputs, performs a weighted summation over its inputs, and applies an activation function to produce an output. The critical distinction from traditional ML is that DL processes raw data without extensive manual preprocessing, instead directly learning hierarchical feature representations through multiple network layers [10] [12]

The most frequently employed DL architectures in healthcare imaging include convolutional neural networks (CNNs) for image analysis, recurrent neural networks (RNNs) and long short-term memory (LSTM) networks for temporal data, and generative adversarial networks (GANs) for synthetic data generation [12] [46] These approaches have demonstrated superior performance in medical imaging classification tasks. Importantly, deep learning algorithms work more accurately on large datasets compared to conventional machine learning algorithms; conversely, machine learning algorithms outperform deep learning in cases of small or medium-sized datasets [10]. This high data requirement creates a critical bottleneck in sensitive healthcare domains where data scarcity, privacy constraints, and institutional fragmentation fundamentally limit centralized data aggregation.

2.2.4 Centralized learning paradigm

The traditional centralized learning paradigm aggregates all available data into a single geographic location—typically a large research hospital, medical center, or data repository—where a unified AI model is trained on the complete dataset [12]. This architecture offers practical advantages: it simplifies model development and validation, enables direct temporal and cross-sectional comparisons of data across diverse patient populations, permits deployment of powerful computational resources at a central facility, and accelerates model training cycles. The centralized approach has dominated healthcare AI research for two decades and has produced the majority of high-performing models cited in clinical literature [12] [46].

However, the centralized learning paradigm faces fundamental limitations rooted in data fragmentation, institutional heterogeneity, and regulatory constraints. Most critically, models trained on data aggregated from a single institution demonstrate significantly degraded performance when deployed in different clinical settings characterized by different patient demographics, imaging hardware specifications, data acquisition protocols, and statistical distributions—a phenomenon widely documented as poor model generalizability [12]. Additionally, centralized architectures require that all sensitive patient data be physically transmitted to, stored within, and processed at a single location, thereby concentrating privacy and security risks at a single point of failure [12]. These limitations directly motivate the development of alternative distributed and privacy-preserving learning paradigms.

2.2.5 Limitations of centralized ML in sensitive and distributed environments

The centralized ML paradigm confronts profound limitations in healthcare environments characterized by stringent data privacy regulations, sensitive patient information, and

geographically dispersed institutional silos.

First, regulatory and legal barriers prohibit centralized data aggregation. Aggregating raw data from personal devices or data centers of multiple organizations to a centralized server poses immediate information leakage risks [32]. Legal restrictions including the Cybersecurity Law of the People's Republic (CLPR) of China, the General Data Protection Regulation (GDPR) in the European Union, the Personal Data Protection Act (PDPA) in Singapore, the California Consumer Privacy Act (CCPA), and the Consumer Privacy Bill of Rights (CPBR) in the United States have been implemented, making data aggregation from distributed devices and multiple regions almost impossible [32]. These regulations enforce data minimization principles—collecting and storing only what is consumer-consented and absolutely necessary [32]. Organizations face legal penalties when transferring protected health information across institutional boundaries [12], rendering centralized data aggregation practically infeasible.

Second, single-center models lack generalizability to diverse populations. AI models trained exclusively on one institution's data often show degraded performance when deployed in different centers with different patient demographics, imaging hardware, and data distributions [12]. Most current studies rely on single-center datasets (ADHD-200, ADNI), lacking large-scale external validation, which limits applicability to real-world clinical practice [46].

Third, data heterogeneity compromises model reproducibility. Small sample sizes, absent external validation, and inconsistencies in data preprocessing compromise reproducibility. High-impact studies frequently report inflated performance estimates due to risks of overfitting and inconsistent acquisition protocols [12].

Fourth, fundamental tension exists between data requirements and privacy constraints. Deep learning requires vast diverse datasets for robust performance, yet healthcare data remains fragmented, siloed, and legally protected at its source [32]. This tension, powerful models require centralized massive datasets, yet data cannot be legally aggregated, motivates alternative approaches. Federated learning enables multiple clients to collaboratively train a model while keeping raw data decentralized without being moved to a single server or data center [32] [36]. Federated learning brings code to the data instead of bringing data to the code, addressing fundamental problems of privacy, ownership, and data locality [32].

2.3 Federated Learning: general concepts

2.3.1 Motivation for federated learning

First introduced by Google in 2016, federated learning (FL) is a distributed machine learning paradigm designed to enable multiple devices or organizations to collaboratively train a shared model without requiring the direct exchange of raw data [50] [34]. Traditional machine learning approaches generally rely on centralized datasets, where all training data is collected and stored in a single location. Although effective, this centralized strategy raises significant concerns regarding privacy, security, and data ownership.

In many domains, particularly healthcare, finance, and mobile applications, data is highly sensitive and cannot easily be transferred across organizations or jurisdictions.

Regulations such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) impose strict constraints on the collection, storage, and sharing of personal information [50]. In addition, institutions are often reluctant to share their data because of legal liabilities, ethical concerns, or competitive interests.

Federated learning was introduced specifically to address these limitations. Instead of transferring raw data to a centralized server, each participant trains the model locally using its own dataset and shares only model updates or learned parameters [50] [34]. A central server then aggregates these local updates to produce a global model that benefits from the collective knowledge of all participants while preserving data privacy [34].

The primary objective of federated learning is therefore to combine collaborative machine learning with privacy preservation [34]. By avoiding centralized data storage, FL reduces the risks associated with data breaches and unauthorized access, enables collaboration among institutions that cannot directly exchange sensitive information, and allows distributed datasets that would otherwise remain isolated to contribute to a common learning process.

2.3.2 FL architecture and workflow

Federated learning generally follows a client-server architecture composed of a central server and multiple distributed clients [50]. The clients may represent hospitals, smartphones, companies, IoT devices, or any other entities possessing local datasets and computational resources. The training process begins when the central server initializes a global model and distributes it to a subset of participating clients. Each client then trains the model locally using its own private dataset through optimization techniques such as stochastic gradient descent (SGD). After local training, the clients send their model updates back to the server instead of transmitting the raw data itself. The server aggregates these updates to generate an improved global model, which is redistributed to the clients for the next training round.

This iterative process continues through multiple communication rounds until the model reaches satisfactory performance or convergence [50]. A fundamental principle of federated learning is that raw data always remains local to the client. Only parameters, gradients, or weight updates are exchanged between the participants and the server. To reduce communication overhead, clients usually perform several local training epochs before synchronization. This design improves communication efficiency while still enabling collaborative learning across distributed datasets.

2.3.3 Roles of server and clients

In federated learning systems, the central server and the participating clients perform different but complementary roles. The server is responsible for coordinating the overall learning process. It initializes the global model, selects participating clients for each communication round, distributes the current global model, collects local updates, and aggregates them into a new global model [50]. The server may also evaluate the global model using validation datasets and monitor the convergence of the training process. The clients, which are also referred to as data owners, are responsible for local model training.

Each client downloads the current global model from the server and trains it using its own private data. After local optimization, the client computes updated model parameters and sends them back to the server. Importantly, the raw local data never leaves the client device or institution. One of the most important characteristics of federated learning is that client datasets are typically non-IID (non-independent and non-identically distributed) [41]. This means that the statistical distributions of the data differ across clients. For example, hospitals located in different regions may treat different patient populations, while smartphone users may generate different behavioral patterns. This heterogeneity makes federated learning fundamentally different from traditional centralized machine learning.

2.3.4 Model aggregation and communication rounds

A key aspect of federated learning is the aggregation of locally trained models into a single global model. The most widely adopted aggregation method is Federated Averaging (FedAvg), introduced by McMahan et al. (2017) [36]. In this approach, each client trains a local version of the model and sends the updated parameters to the central server. The server then computes a weighted average of these local updates according to the size of each client's dataset.

$$w^{t+1} = \sum_{k=1}^{K} \frac{n_k}{n} w_k^{t+1} \tag{2.1}$$

In Equation 2.1, K represents the number of participating clients, n_k corresponds to the number of samples stored on client k, and n represents the total number of samples across all clients. The aggregation process gives greater importance to clients possessing larger datasets [36]. Federated learning proceeds through repeated communication rounds. During each round, the server distributes the current global model to the clients, local training is performed independently, and the resulting updates are transmitted back to the server for aggregation. Communication efficiency represents one of the main challenges in federated learning because transmitting large neural network parameters repeatedly can require substantial bandwidth [50]. To reduce communication costs, clients often perform multiple local training epochs before sending updates. This introduces a trade-off between local computation and communication frequency. Increasing local computation decreases communication overhead but may also increase the divergence between local and global models.

2.3.5 Types of Federated Learning

Federated learning can be categorized according to the distribution of data samples and feature spaces across the participating clients [50] [41]. As illustrated in Figure 2.1, three principal federated learning paradigms can be distinguished: Horizontal Federated Learning (HFL), Vertical Federated Learning (VFL), and Federated Transfer Learning (FTL). Each paradigm addresses a different data distribution scenario and is therefore suitable for different collaborative learning applications.

Horizontal Federated Learning (HFL). Horizontal Federated Learning is the most widely adopted federated learning paradigm. In this setting, participating clients

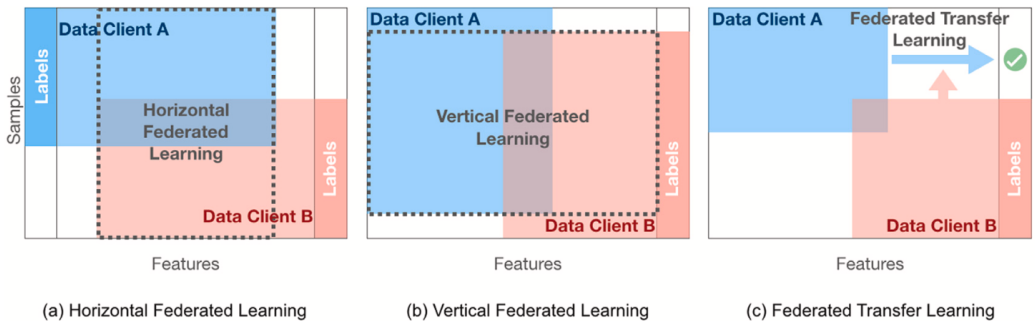


Figure 2.1: Classification of federated learning according to the distribution of samples and feature spaces across participating clients. Adapted from [41].

share the same feature space and label space but possess different data samples, as defined in Equation 2.2. A typical example is a group of hospitals that collect the same clinical variables while treating different patient populations. By collaboratively training a common model, institutions can benefit from a larger effective training dataset without exchanging their raw patient records.

$$X_i = X_j, \quad Y_i = Y_j, \quad I_i \neq I_j \tag{2.2}$$

where X denotes the feature space, Y the label space, and I the set of data samples owned by each client.

Vertical Federated Learning (VFL). In Vertical Federated Learning, participating organizations share information about the same individuals but possess different sets of features. This scenario is represented by Equation 2.3. For example, one healthcare institution may store laboratory test results while another maintains medical imaging data for the same group of patients. Since the data correspond to the same individuals, VFL requires an initial entity-matching phase to align records before collaborative training can be performed.

$$X_i \neq X_j, \quad Y_i = Y_j, \quad I_i = I_j \tag{2.3}$$

Federated Transfer Learning (FTL). Federated Transfer Learning addresses the most heterogeneous scenario, in which participating clients share neither the same feature space nor the same data samples, as shown in Equation 2.4. Because the available data differ substantially across organizations, conventional federated optimization is insufficient. Instead, transfer learning techniques are employed to transfer knowledge between related tasks or domains, enabling collaboration even when the participating datasets exhibit minimal overlap.

$$X_i \neq X_j, \quad Y_i \neq Y_j, \quad I_i \neq I_j \tag{2.4}$$

Among these paradigms, Horizontal Federated Learning is by far the most common in healthcare because different hospitals generally collect the same types of clinical information while treating different groups of patients. Consequently, the framework proposed in this thesis adopts the Horizontal Federated Learning setting.

2.3.6 Existing frameworks for federated learning

Several open-source frameworks have been developed to support federated learning research and deployment [50]. Among the most widely used platforms is FATE (Federated AI Technology Enabler), an industrial-grade framework supporting horizontal, vertical, and transfer federated learning. FATE also integrates privacy-preserving mechanisms such as homomorphic encryption and differential privacy.

TensorFlow Federated (TFF), developed by Google, is built on TensorFlow and provides tools for simulating and implementing federated learning algorithms, including Federated Averaging. Another important framework is PySyft, which integrates with PyTorch and TensorFlow to provide federated learning capabilities together with secure computation and privacy-preserving techniques. PaddleFL, developed by Baidu, offers

additional support for federated learning and focuses on usability and integration with the PaddlePaddle deep learning ecosystem [41].

In addition to these, Flower is a comprehensive framework that enables large-scale federated learning experiments, capable of scaling to workloads of up to 15 million clients. It is particularly distinguished by its ability to seamlessly transition from experimental simulations to execution on real, highly heterogeneous edge devices. Furthermore, Flower is designed to be language-agnostic and machine learning framework-agnostic [6]. This allows researchers to integrate their preferred ML pipelines and tools—such as PyTorch or TensorFlow—without being locked into a single ecosystem.

Another notable platform is FedML, which serves as an open research library and benchmark designed to facilitate both algorithmic innovation and fair performance comparisons. FedML is highly versatile, supporting three distinct computing paradigms: single-machine simulation, distributed computing, and on-device training for Internet of Things (IoT) and mobile edge devices. A core feature of FedML is its worker-oriented programming interface [24]. This interface provides the flexibility needed to support diverse network topologies beyond traditional centralized models, easily accommodating decentralized federated learning, vertical federated learning, and split learning configurations.

Although these frameworks simplify implementation, practical federated learning systems still face important challenges related to privacy protection, communication efficiency, non-IID data distributions, and security against malicious attacks.

2.4 Federated Learning in healthcare

2.4.1 Why FL is suitable in healthcare

Federated learning (FL) is considered particularly suitable for healthcare because medical data is highly sensitive, distributed across many institutions, and difficult to share due to legal and ethical restrictions [23] [40]. Modern artificial intelligence systems, especially deep learning models, require large quantities of data in order to achieve reliable performance. However, in healthcare, a single hospital often possesses only a limited number of patient records or medical images related to a specific disease. This limitation is especially evident for rare diseases and specialized diagnostic tasks [23]. Traditionally, machine learning models are trained using centralized approaches in which all datasets are collected and stored in a single location. In healthcare, this strategy creates major problems because patient information is protected by strict regulations such as the General Data Protection Regulation (GDPR) in Europe and HIPAA in the United States. These regulations strongly limit the exchange of medical data across institutions and require healthcare providers to guarantee patient confidentiality and security [23]. Beyond legal constraints, centralized storage also increases cybersecurity and privacy risks because large amounts of sensitive patient information are concentrated in one repository. Hospitals are therefore often unwilling to transfer raw data to external organizations or research centers [40]. Federated learning addresses these limitations by allowing institutions to collaboratively train machine learning models without sharing raw patient data. Instead of transferring datasets, each institution trains the model locally and shares

only model updates or parameters with a central server [23]. In this way, the data always remains inside the hospital infrastructure while the global model benefits from the knowledge of multiple institutions. This collaborative approach improves privacy protection, facilitates regulatory compliance, and enables the creation of larger and more diverse training datasets. As a result, federated learning has become one of the most promising paradigms for healthcare artificial intelligence [40].

2.4.2 FL across hospitals and clinical institutions

In healthcare federated learning systems, hospitals and clinical institutions collaborate to train a common machine learning model while maintaining local control over their data [23]. The architecture generally follows a client-server structure in which a central server coordinates the learning process and hospitals act as distributed clients. The training process begins when the server initializes a global model and distributes it to the participating hospitals. Each institution trains the model locally using its own patient data, such as medical images, laboratory results, or electronic health records. Once local training is completed, the hospitals send the updated model parameters back to the server rather than sending the underlying patient data. The server aggregates the received updates, usually through weighted averaging, and produces a new global model that is redistributed to the participants [23]. This procedure is repeated over multiple communication rounds until the model converges. An important characteristic of healthcare federated learning is the heterogeneity of medical data across institutions. Hospitals may use different imaging devices, acquisition protocols, clinical workflows, and electronic record systems. In addition, patient populations may vary significantly in terms of demographics, disease prevalence, and comorbidities [23]. Consequently, the local datasets are often non-IID, meaning that their statistical distributions differ from one another. Federated learning attempts to address this problem by enabling collaborative training while preserving institutional independence. This is particularly important in healthcare because models trained using data from a single hospital often fail to generalize well when applied to patients from other clinical environments. FL therefore allows institutions to benefit from shared knowledge without losing ownership or control of their sensitive medical data [40].

2.4.3 Healthcare applications: diagnosis, imaging, prediction and monitoring

Federated learning has been applied to a wide range of healthcare applications, especially in medical imaging, predictive analytics, and patient monitoring. One of the most active research areas is medical image analysis. Hospitals can collaboratively train deep learning models for disease detection and diagnosis while preserving patient privacy [23].

Several studies have explored federated approaches for brain tumor segmentation using MRI scans, pneumonia detection from chest X-rays, intracranial hemorrhage detection from CT scans, and COVID-19 diagnosis from medical images [40]. These applications benefit greatly from distributed learning because diagnostic accuracy improves

when models are exposed to larger and more diverse datasets collected from multiple institutions. Federated learning is also widely used with electronic health records (EHRs). In this context, collaborative predictive models can support mortality prediction, disease progression analysis, patient deterioration prediction, and hospital length-of-stay estimation [23]. Since healthcare systems differ across hospitals, federated learning helps develop models that are more robust and generalizable across different clinical environments.

Another important application area involves wearable devices and Internet-of-Medical-Things (IoMT) technologies. Smart sensors and wearable devices continuously generate physiological information such as heart rate, blood pressure, physical activity, and sleep quality [40]. Federated learning allows these distributed devices to contribute to global predictive models without requiring users to share their raw personal data. FL has also attracted attention in genomic and multi-modal healthcare research. In some scenarios, different institutions possess complementary information related to the same patients, such as imaging data, genomic profiles, and clinical records. Federated learning enables collaborative analysis of these heterogeneous data sources and supports applications such as biomarker discovery, disease subtype identification, and personalized medicine [23].

2.4.4 Challenges in healthcare FL: heterogeneity, bias, communication, validation

Despite its advantages, federated learning in healthcare still faces several important challenges. One of the main difficulties is data heterogeneity. Medical datasets differ substantially across hospitals because of variations in imaging equipment, acquisition settings, patient demographics, and clinical procedures [23]. These differences can negatively affect model convergence and reduce the performance of the global model. Another challenge concerns the limited availability of labeled medical data. Medical annotation is expensive and often requires expert clinicians.

As a result, many hospitals possess relatively small datasets, especially for rare diseases [23]. In addition, healthcare datasets are frequently imbalanced because healthy cases are more common than pathological ones. This imbalance can introduce bias into the learning process and reduce the accuracy of predictions for underrepresented conditions. Communication efficiency also represents a significant issue. Federated learning requires repeated exchanges of model updates between hospitals and the central server. Modern deep learning models may contain millions of parameters, which creates substantial communication overhead, particularly in large federated networks [40].

Although federated learning improves privacy protection compared to centralized learning, it is not completely immune to security threats. Research has shown that sensitive information may sometimes be reconstructed from gradients or shared parameters [23]. For this reason, additional techniques such as differential privacy, secure aggregation, and homomorphic encryption are often integrated into FL systems. Bias and fairness are also major concerns in healthcare AI. If certain demographic groups or disease categories are poorly represented in the training data, the resulting model may perform unequally across populations [23]. Ensuring fairness and reducing bias are therefore essential requirements for safe and ethical deployment. Finally, model validation and generalization remain open problems. Many studies evaluate federated models

only on institutions participating in the training process. However, clinical deployment requires models that can generalize effectively to completely new hospitals and patient populations [23].

2.4.5 Evaluation, deployment, and clinical integration issues

The deployment of federated learning systems in healthcare involves not only technical considerations but also organizational, clinical, and regulatory challenges. A central issue concerns model evaluation. In healthcare applications, predictive accuracy alone is insufficient because models must also demonstrate robustness, reliability, fairness, and clinical usefulness [23]. Proper validation generally requires external evaluation using datasets from institutions that were not involved in the training process. Without external validation, model performance may appear artificially optimistic.

Clinical integration is another important challenge. Federated learning systems must interact with existing hospital infrastructures, including electronic health record systems, imaging platforms, and cybersecurity frameworks [40]. Since hospitals often use different technologies and workflows, integrating FL systems into clinical practice can be complex and resource-intensive. Regulatory approval also represents a significant barrier. Healthcare AI systems may require authorization from agencies such as the FDA in the United States or the EMA in Europe. However, regulatory frameworks for federated learning are still evolving, particularly because federated models may continue to change over time through distributed updates [23].

In addition, successful healthcare federations require strong governance structures. Participating institutions must define common rules regarding data quality, participation procedures, cybersecurity policies, intellectual property ownership, and responsibilities in case of system failures [23]. In many cases, organizational coordination and governance become more challenging than the technical implementation itself. Finally, clinical adoption depends heavily on trust from healthcare professionals. Doctors and clinicians must understand the capabilities and limitations of AI systems and retain the ability to question or override automated predictions when necessary. Consequently, explainability, transparency, and human supervision remain fundamental requirements for the integration of federated learning into real clinical practice.

2.5 Legal and regulatory aspects for AI and Federated Learning in healthcare

2.5.1 Sensitive Health Data and Legal Obligations

Health data are among the most sensitive categories of personal information and are therefore subject to stringent legal protection worldwide [29]. Medical records reveal detailed information about an individual's physical and mental health, making unauthorized access or misuse particularly harmful to both patients and healthcare providers [59]. Modern privacy regulations broadly define health data as any personal information relating to an individual's past, present, or future physical or mental condition that is collected during healthcare provision [59].

The widespread digitalization of healthcare has significantly increased both the availability and the sensitivity of medical information. Electronic Health Records (EHRs), telemedicine platforms, wearable devices, Internet of Medical Things (IoMT) technologies, and remote monitoring systems continuously generate large volumes of patient data. At the same time, advanced AI and machine learning techniques can infer sensitive information beyond what is explicitly recorded, creating additional privacy risks and reducing patients' control over how their data are processed [59].

Consequently, healthcare organizations are subject to comprehensive legal obligations aimed at protecting patient privacy and ensuring secure data management. Hospitals, research institutions, and other healthcare stakeholders must implement appropriate technical and organizational safeguards to prevent unauthorized access, data breaches, and misuse of sensitive information. Failure to adequately protect health data may compromise patient safety, undermine public trust, and expose organizations to significant legal and regulatory consequences [29].

2.5.2 GDPR and Data Protection Requirements

The General Data Protection Regulation (GDPR), adopted in 2016 and enforced since 2018, represents one of the most comprehensive legal frameworks for the protection of personal data worldwide [2] [26]. Unlike traditional privacy regulations primarily focused on confidentiality, the GDPR adopts a broader governance perspective based on fairness, transparency, accountability, and lawful data processing. These principles are particularly relevant in healthcare, where medical information is classified as highly sensitive personal data.

Several GDPR principles directly influence the design of Federated Learning systems, including data minimization, purpose limitation, proportionality, and fair processing [59]. Since Federated Learning keeps patient data locally stored and exchanges only model updates, it naturally supports the principle of data minimization. Nevertheless, this decentralized architecture does not automatically guarantee GDPR compliance. Recent studies have shown that model parameters and gradients may still reveal sensitive information through attacks such as Membership Inference and Gradient Inversion, meaning that Federated Learning remains subject to GDPR obligations [45].

The GDPR also strengthens the rights of data subjects by requiring informed consent and providing individuals with rights including data access, rectification, erasure, portability, and objection to processing [26] [59]. In Federated Learning environments, obtaining meaningful informed consent becomes more challenging because patients must understand not only the medical application, but also the collaborative training process, the participating institutions, and the residual privacy risks associated with exchanging model updates.

Another important requirement concerns Data Protection Impact Assessments (DPIAs), which are mandatory when processing sensitive healthcare information [59]. While centralized systems typically assign accountability to a single organization, Federated Learning distributes responsibilities across multiple participants, making governance and risk assessment more complex. Similar challenges arise when determining whether exchanged

model updates should themselves be considered personal data, since recent research suggests that they may leak information about individual training samples [45].

Finally, the GDPR promotes the principle of *privacy by design*, requiring privacy and security mechanisms to be integrated into AI systems from the earliest stages of development [59]. Within Federated Learning, this principle motivates the adoption of additional privacy-preserving techniques such as Differential Privacy, Secure Aggregation, and Homomorphic Encryption to reduce the risk of information leakage [45]. Although these mechanisms improve privacy protection and support regulatory compliance, they also introduce trade-offs involving model utility, computational complexity, and communication overhead. Consequently, healthcare organizations must carefully balance legal compliance with the practical performance of Federated Learning systems, particularly given the substantial financial penalties associated with GDPR violations [59].

2.5.3 HIPAA and Healthcare Data Governance

In the United States, the Health Insurance Portability and Accountability Act (HIPAA), enacted in 1996, constitutes the primary federal framework governing the privacy and security of Protected Health Information (PHI). HIPAA establishes requirements for healthcare providers, insurers, and related organizations regarding the collection, storage, processing, and disclosure of patient information. The regulation is primarily composed of three complementary rules: the Privacy Rule, which defines patient rights and limits the use and disclosure of PHI; the Security Rule, which specifies the administrative, technical, and physical safeguards required to protect electronic health information; and the Breach Notification Rule, which mandates the reporting of unauthorized disclosures of unsecured PHI to affected individuals and regulatory authorities [29].

To comply with HIPAA, healthcare organizations must implement documented security policies, perform regular risk assessments, maintain audit mechanisms, and provide workforce training. However, unlike the GDPR, HIPAA applies primarily to covered entities and their business associates, resulting in a narrower regulatory scope that does not explicitly address many emerging AI technologies, including the Internet of Medical Things (IoMT) and Federated Learning systems [29].

Although Federated Learning inherently supports data protection by keeping patient information local to each institution, organizations must still ensure secure communication, effective governance, and compliance with existing healthcare regulations [27]. The absence of specific HIPAA provisions for distributed AI systems therefore creates uncertainty regarding accountability, liability, and security responsibilities when multiple organizations collaboratively train machine learning models.

2.5.4 AI-specific Risks in Healthcare

The adoption of Artificial Intelligence in healthcare introduces challenges that extend beyond those associated with traditional information systems [29]. Because AI models directly influence clinical decision-making, errors or vulnerabilities may have significant consequences for patient safety, privacy, and trust. Consequently, modern healthcare AI systems must address not only predictive performance but also fairness, transparency,

robustness, and accountability.

One of the principal concerns is algorithmic bias. Machine learning models trained on non-representative datasets may systematically underperform for underrepresented demographic groups, potentially leading to inaccurate diagnoses, inappropriate treatments, and increased healthcare disparities [29]. Ensuring the availability of diverse and representative training data is therefore essential for developing equitable AI systems.

Another major challenge concerns accountability and interpretability. When AI-assisted clinical decisions produce adverse outcomes, determining responsibility may involve multiple stakeholders, including healthcare providers, software developers, and data contributors [29]. At the same time, many deep learning models operate as "black boxes", generating predictions without providing transparent explanations. This lack of interpretability can reduce clinicians' confidence in AI-assisted recommendations and complicate informed decision-making.

Healthcare AI systems are also exposed to significant security and privacy threats. Adversaries may exploit vulnerabilities through attacks such as Membership Inference or Gradient Inversion to recover sensitive information from trained models, even when raw patient data are not directly shared [29]. In addition, adversarial examples—small, carefully designed perturbations to input data—can cause machine learning models to produce incorrect predictions, potentially leading to unsafe clinical decisions.

These risks highlight that the deployment of AI in healthcare requires more than high predictive accuracy. Robust security mechanisms, privacy-preserving techniques, explainable models, and well-defined governance frameworks are essential to ensure that AI systems remain trustworthy, reliable, and safe for clinical use.

2.5.5 The EU AI Act and High-risk AI Systems

The European Union Artificial Intelligence Act (EU AI Act), formally adopted in 2024 and progressively entering into force between 2025 and 2026, establishes the first comprehensive risk-based regulatory framework for artificial intelligence within the European Union [29]. The regulation classifies AI systems according to their potential societal impact, with healthcare applications used for diagnosis, treatment planning, patient monitoring, or other clinically relevant decisions explicitly categorized as *high-risk*. Consequently, these systems must satisfy stringent requirements concerning safety, robustness, transparency, and human oversight before deployment.

To comply with the AI Act, high-risk AI systems must undergo conformity assessments and maintain extensive technical documentation, quality management procedures, and continuous post-market monitoring [29]. Developers are also required to document model architectures, training methodologies, dataset characteristics, known limitations, and performance across different demographic groups. In addition, the regulation emphasizes the use of representative datasets and the implementation of meaningful human oversight to reduce bias and ensure that clinicians retain control over AI-assisted decisions [29].

These requirements introduce important challenges for Federated Learning in healthcare. Unlike centralized machine learning, Federated Learning relies on distributed datasets collected by multiple institutions, each characterized by different acquisition

protocols, annotation practices, and patient populations. Such heterogeneity complicates compliance with regulatory requirements concerning data governance, representativeness, robustness, and quality assurance [29]. Furthermore, because regulators require evidence regarding dataset quality without necessarily permitting access to the underlying patient data, balancing regulatory verification with privacy preservation remains a significant challenge [51].

The distributed nature of Federated Learning also complicates accountability, technical documentation, and lifecycle management. Responsibilities for monitoring system performance, identifying vulnerabilities, and maintaining regulatory compliance are shared across multiple organizations, making governance substantially more complex than in centralized AI systems [29]. Similarly, decentralized training reduces the traceability of model updates and data provenance, creating additional challenges for auditing and quality assurance.

Despite these challenges, Federated Learning aligns closely with several objectives of the EU AI Act by supporting privacy-preserving and decentralized data processing [51]. Keeping patient data locally stored naturally supports privacy-by-design principles while reducing risks associated with centralized data collection. However, additional mechanisms such as Differential Privacy, Secure Aggregation, and Homomorphic Encryption are often required to further strengthen privacy and regulatory compliance, introducing trade-offs between security, computational cost, and model performance. Consequently, future Federated Learning systems must satisfy not only technical performance requirements but also the legal and ethical obligations established by modern AI regulation.

2.5.6 Transparency, Accountability and Human Oversight

Modern regulatory frameworks increasingly emphasize transparency, accountability, and human oversight as fundamental principles for the safe deployment of Artificial Intelligence in healthcare [29]. Beyond achieving high predictive performance, AI systems must remain understandable, controllable, and legally accountable, particularly when they contribute to clinical decision-making.

Transparency requires healthcare organizations to clearly communicate when AI systems are involved in patient care and to maintain comprehensive documentation covering the entire system lifecycle, including model development, training, validation, deployment, and monitoring [29] [59]. It also involves reporting known limitations, potential sources of bias, and variations in model performance across different patient populations. Such documentation supports both regulatory compliance and the clinical validation of AI-assisted decisions.

Accountability ensures that responsibility for AI-assisted decisions remains clearly defined. Regulations such as the GDPR establish that healthcare organizations, acting as data controllers, retain responsibility for the outcomes of automated processing, even when external software vendors or third-party developers contribute to the implementation of AI technologies [26]. Consequently, clear governance mechanisms are required to allocate responsibilities in cases involving incorrect predictions, biased outcomes, or system failures.

Human oversight represents another essential principle of trustworthy AI. Rather than

replacing healthcare professionals, AI systems should support clinical decision-making while allowing clinicians to interpret, validate, and, when necessary, override automated recommendations [26] [29]. This *human-in-the-loop* approach preserves clinical autonomy and ensures that final medical decisions remain under professional supervision.

To support these principles, healthcare organizations should implement continuous monitoring, regular security assessments, and comprehensive audit trails throughout the operational lifecycle of AI systems [29]. Together, transparency, accountability, and human oversight establish the governance foundation required for trustworthy and legally compliant AI deployment in healthcare.

2.6 Privacy and Security Challenges in Federated Learning

Federated learning significantly reduces the need to centralize sensitive data by enabling collaborative model training across multiple institutions while keeping data locally stored. Despite this advantage, federated learning remains vulnerable to a variety of privacy and security threats that may compromise the confidentiality of the training data, the integrity of the learned model, or the reliability of the collaborative learning process.

Unlike centralized machine learning, the iterative exchange of model updates introduces new attack surfaces that can be exploited by honest-but-curious or malicious participants. Adversaries may attempt to infer sensitive information about the local training data, reconstruct original samples from exchanged gradients, or manipulate the global model by injecting poisoned updates. Consequently, protecting federated learning systems requires not only preserving data confidentiality but also ensuring the robustness of the aggregation process against adversarial behaviour.

Several defence mechanisms have been proposed to mitigate these threats. Differential Privacy reduces information leakage by perturbing model updates before transmission, Homomorphic Encryption protects exchanged parameters during communication and aggregation, while robust aggregation strategies aim to limit the influence of malicious client updates. These complementary techniques form the basis of many privacy-preserving federated learning frameworks and are therefore examined in the remainder of this section.

2.6.1 Cybersecurity perspective and Threat Model

Although federated learning enhances privacy by keeping training data locally stored, it also introduces new security challenges arising from the distributed nature of the learning process. Unlike centralized machine learning, where training is performed within a single trusted environment, federated learning requires the repeated exchange of model updates between multiple clients and a central server. This communication process creates several attack surfaces that may be exploited by adversaries to compromise the confidentiality, integrity, or availability of the collaborative learning system.

Figure 2.2 illustrates the principal attack surfaces in a typical federated learning environment. Adversaries may compromise participating clients, intercept model updates during communication, or target the aggregation server to infer sensitive information, manipulate the global model, or disrupt the training process. Consequently, federated learning systems must address a broader range of security threats than conventional

centralized learning approaches. The security threats affecting federated learning are

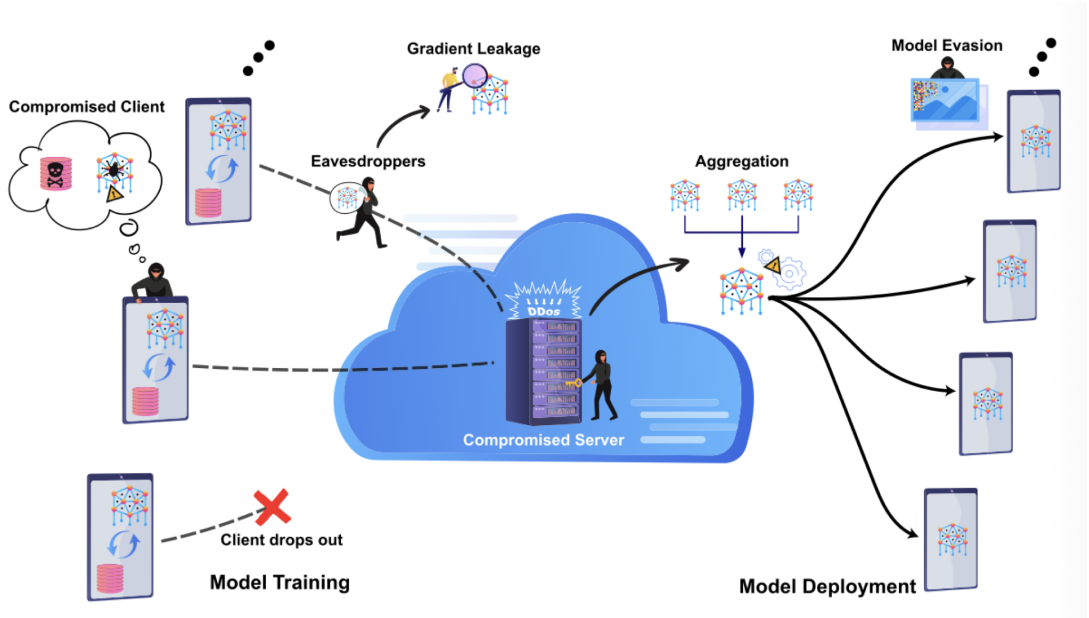


Figure 2.2: Attack surfaces and adversarial threats in a federated learning system. Potential adversaries may compromise clients, intercept model updates during communication, or attack the aggregation server, leading to privacy leakage, model poisoning, or service disruption. Source: [8]

commonly classified according to the security property they target:

- **Confidentiality attacks** aim to infer sensitive information about the local training data without directly accessing the datasets. These attacks exploit the information contained in exchanged gradients, model updates, or the final global model to recover private information about participating clients.
- **Integrity attacks** seek to manipulate the collaborative training process by injecting malicious data or model updates. Their objective is to influence the behaviour of the global model while remaining undetected during aggregation.
- **Availability attacks** attempt to disrupt or degrade the federated learning process by preventing successful communication, reducing model convergence, or denying access to participating clients. Examples include denial-of-service attacks and client dropout scenarios.

Although all three categories represent important security concerns, this thesis focuses on threats targeting *confidentiality* and *integrity*. Specifically, the security evaluation considers Membership Inference Attacks and Gradient Inversion Attacks as representative privacy attacks, together with the Frequency-Injection Backdoor Attack as a representative integrity attack. These attacks were selected because they represent realistic adversarial scenarios in healthcare federated learning and directly evaluate the effectiveness of

the privacy-preserving mechanisms proposed in this work.

2.6.2 Privacy Attacks

Privacy attacks exploit the information exchanged during the federated learning process to infer sensitive knowledge about the participants' local datasets. Although federated learning prevents the direct sharing of raw data, the iterative communication of model parameters or gradients may unintentionally leak information regarding the underlying training samples. Several studies have demonstrated that adversaries can exploit this information to infer membership, reconstruct training data, or recover sensitive attributes without compromising the communication protocol itself [31] [38] [60].

Among the various privacy attacks proposed in the literature, Membership Inference Attacks and Gradient Inversion Attacks have received particular attention due to their effectiveness and their direct impact on the confidentiality of sensitive medical data. These attacks target complementary aspects of privacy leakage and therefore provide a comprehensive evaluation of the security guarantees offered by federated learning systems.

Membership Inference Attacks

Membership Inference Attacks (MIAs) aim to determine whether a particular data sample was included in the training dataset of a machine learning model. Although the attack does not directly reconstruct the original data, revealing the participation of an individual in the training process may itself disclose sensitive information. In healthcare applications, for example, confirming that a patient's medical record contributed to the training of a disease-specific model may inadvertently reveal information about the patient's medical condition or association with a particular healthcare institution. Consequently, Membership Inference Attacks have become one of the most widely studied privacy threats in machine learning and federated learning [31] [38].

The first practical Membership Inference Attack was introduced by Shokri et al. [43], who proposed a shadow-model framework capable of learning the behavioural differences between training samples (members) and previously unseen samples (non-members). Their approach trains multiple shadow models to imitate the target model and subsequently employs a dedicated attack classifier to distinguish members from non-members based on prediction confidence. This work demonstrated that deep neural networks often memorize characteristics of the training data, making successful membership inference possible even without direct access to the original dataset.

Building upon this work, Yeom et al. [55] showed that effective membership inference can also be achieved through considerably simpler strategies. Rather than relying on computationally expensive shadow models, their approach exploits the observation that training samples generally exhibit lower prediction loss than unseen samples. This finding established a strong relationship between model overfitting and privacy leakage, demonstrating that highly accurate models may simultaneously expose more information about their training data.

More recent research has focused on improving both the scalability and practicality of Membership Inference Attacks. Watson et al. [47] introduced the Quantile Membership

Inference Attack (Quantile MIA), which replaces shadow-model training with quantile regression models learned from publicly available data. By estimating confidence thresholds directly from prediction score distributions, Quantile MIA significantly reduces computational complexity while maintaining competitive attack performance. This methodology has become particularly attractive for modern deep learning models due to its scalability and reduced assumptions regarding the attacker's knowledge.

In federated learning, Membership Inference Attacks can exploit both the exchanged model updates and the final global model to infer the participation of individual training samples. Since healthcare datasets frequently contain highly sensitive patient information, even identifying whether a patient's record contributed to model training may constitute a serious privacy violation. Consequently, evaluating resistance against Membership Inference Attacks has become a standard component of security assessments for privacy-preserving federated learning systems.

Despite significant progress, several challenges remain. The effectiveness of Membership Inference Attacks depends on multiple factors, including model architecture, dataset complexity, client data distribution, and model overfitting. Current research therefore focuses not only on developing stronger attacks but also on designing defence mechanisms capable of reducing membership leakage while preserving predictive performance. Among these countermeasures, Differential Privacy has emerged as one of the most effective approaches and is discussed later in this section.

Gradient Inversion Attacks

Gradient Inversion Attacks (GIAs) constitute another major category of privacy attacks against federated learning. Unlike Membership Inference Attacks, which determine whether a sample participated in the training process, Gradient Inversion Attacks attempt to reconstruct the original training data directly from the gradients or model updates exchanged during collaborative optimization. Since gradients are computed from both the model parameters and the input data, they inevitably contain information about the training samples, which can be exploited by an adversary to recover sensitive information [31] [38].

The first practical demonstration of this vulnerability was presented by Zhu et al. [61], who introduced the *Deep Leakage from Gradients* (DLG) attack. DLG formulates data reconstruction as an optimization problem in which randomly initialized dummy inputs and labels are iteratively optimized until the gradients produced by the dummy data closely match the gradients observed by the attacker. Their results demonstrated that surprisingly accurate image reconstructions could be obtained from a single gradient update, revealing that gradient sharing itself may pose a significant privacy risk.

Although DLG established the feasibility of gradient reconstruction, its optimization procedure often suffered from slow convergence and limited reconstruction quality, particularly when applied to deeper neural networks or larger datasets. To overcome these limitations, Geiping et al. [19] proposed the *Inverting Gradients* attack, which introduced several important improvements, including cosine-similarity gradient matching, total variation regularization, and a more robust optimization strategy. These enhancements significantly improved reconstruction fidelity and enabled successful attacks against

modern convolutional neural networks trained on complex image datasets, demonstrating that gradient leakage remains a serious concern even for large-scale deep learning models.

Further improvements have continued to increase the effectiveness of gradient inversion attacks. Yin et al. [57] proposed GradInversion, which incorporates image priors and additional regularization techniques to recover batches of high-quality images from shared gradients. More recently, several studies have investigated the influence of training hyperparameters on reconstruction quality, showing that factors such as batch size, network architecture, optimization algorithm, and gradient sparsity substantially affect the amount of information that can be recovered from exchanged gradients [57] [28].

The effectiveness of Gradient Inversion Attacks has important implications for healthcare federated learning. Even though patient records remain stored locally, successful reconstruction of medical images or other sensitive clinical data from communicated gradients may violate patient confidentiality and compromise compliance with privacy regulations. Consequently, protecting model updates has become a fundamental requirement for secure federated learning systems.

Current research therefore focuses on developing defence mechanisms capable of reducing gradient leakage while preserving model utility. Differential Privacy, secure aggregation, gradient compression, and Homomorphic Encryption have all been investigated as potential countermeasures, each introducing different trade-offs between privacy protection, computational overhead, and predictive performance [38] [60]. Among these approaches, Differential Privacy has emerged as one of the most widely adopted techniques due to its formal privacy guarantees and compatibility with existing federated learning frameworks.

2.6.3 Integrity Attacks

While privacy attacks seek to infer confidential information from the federated learning process, integrity attacks aim to manipulate the collaborative training procedure in order to alter the behaviour of the global model. Rather than extracting sensitive information, the adversary attempts to influence the learning process by injecting malicious data or model updates, ultimately compromising the correctness and reliability of the trained model.

Among the various integrity attacks proposed in the literature, backdoor attacks have emerged as one of the most effective and challenging threats against federated learning systems. These attacks exploit the decentralized nature of collaborative learning by allowing one or more compromised clients to poison their local training process while remaining indistinguishable from honest participants. Since the server typically has no access to the clients' private datasets, detecting malicious updates becomes particularly difficult, making backdoor attacks an important research topic in secure federated learning.

Backdoor Attacks

Backdoor attacks are a class of data poisoning attacks in which an adversary intentionally modifies a subset of the training data by embedding a predefined trigger and assigning the manipulated samples to a target label. During training, the model learns to associate the

trigger with the attacker's chosen class while maintaining normal predictive performance on benign inputs. As a result, the backdoor remains dormant during standard operation and is activated only when the trigger is present in the input data [22].

In federated learning, backdoor attacks pose a particularly severe threat because each participating client performs local training independently. A malicious client can therefore poison its private dataset or manipulate its local model updates before transmitting them to the central server. Since the server aggregates updates without direct access to the local training data, malicious contributions may be incorporated into the global model without being detected [3] [52].

Early studies focused primarily on visible spatial triggers, such as small image patches or geometric patterns embedded into the input images. Gu et al. [22] demonstrated the feasibility of such attacks through the BadNets framework, showing that deep neural networks can successfully learn hidden backdoor behaviours while maintaining high classification accuracy on clean samples. Although effective, these visible triggers are often perceptible and may therefore be detected through visual inspection or data sanitization techniques.

To improve stealthiness, more recent research has explored imperceptible trigger generation techniques. Among these approaches, Feng et al. [16] proposed the *Frequency-Injection Backdoor Attack* (FIBA), which embeds the malicious trigger in the frequency domain by manipulating the Fourier spectrum of the input images. Unlike conventional spatial-domain triggers, frequency-domain perturbations produce minimal visual artifacts while preserving a high attack success rate, making them significantly more difficult to detect using conventional defence mechanisms.

The increasing sophistication of backdoor attacks has motivated the development of robust aggregation algorithms capable of identifying and mitigating malicious client updates during federated training. Approaches based on Byzantine-robust aggregation, anomaly detection, and kernel similarity analysis have demonstrated promising results in limiting the influence of poisoned updates while preserving the contributions of honest participants [7] [62]. Nevertheless, designing aggregation mechanisms that simultaneously achieve robustness, scalability, and high predictive performance remains an open research challenge.

Because of their effectiveness and practical relevance, backdoor attacks have become one of the standard benchmarks for evaluating the robustness of federated learning systems. In particular, frequency-domain attacks such as FIBA represent one of the most challenging poisoning strategies currently reported in the literature due to their high stealthiness and strong attack performance.

2.6.4 Privacy-Preserving Mechanisms

The increasing sophistication of privacy and integrity attacks has demonstrated that federated learning alone cannot guarantee the confidentiality of training data or the robustness of the collaborative learning process. Although keeping data locally stored significantly reduces the risks associated with centralized data collection, information may still be inferred from exchanged model updates, while malicious participants may manipulate the global model through poisoned contributions.

To address these challenges, numerous privacy-preserving mechanisms have been proposed in the literature. These mechanisms aim to protect different stages of the federated learning pipeline by limiting information leakage, securing model communication, or improving the robustness of the aggregation process. Among the various solutions, Differential Privacy, Homomorphic Encryption and robust aggregation algorithms have emerged as the most widely adopted approaches due to their effectiveness and compatibility with existing federated learning frameworks [38] [31] [60].

The following subsections review these techniques, discussing their operating principles, advantages, limitations, and their role in enhancing the privacy and security of federated learning systems.

Differential Privacy

Differential Privacy (DP) is one of the most widely adopted techniques for protecting sensitive information during machine learning. Originally introduced by Dwork et al. [13], Differential Privacy provides a rigorous mathematical definition of privacy by ensuring that the output of a computation remains statistically similar regardless of whether any individual sample is included in the input dataset. Consequently, an adversary observing the released information cannot reliably determine whether a particular individual's data contributed to the computation.

Formally, a randomized mechanism $\mathcal{M}$ satisfies (ε, δ)-Differential Privacy if, for any pair of neighbouring datasets differing by a single record and for every measurable output set S, the following condition holds:

$$\Pr[\mathcal{M}(D) \in S] \leq e^{\varepsilon} \Pr[\mathcal{M}(D') \in S] + \delta, \tag{2.5}$$

As defined in Equation 2.5, ε denotes the privacy budget and δ represents a small probability of violating the privacy guarantee. Smaller values of ε correspond to stronger privacy protection, whereas larger values generally preserve higher model utility.

In federated learning, Differential Privacy is typically implemented by clipping the gradients or model updates computed during local training and subsequently adding carefully calibrated random noise before transmission to the aggregation server [21] [37]. By perturbing the exchanged information, Differential Privacy reduces the amount of sensitive information available to potential adversaries, thereby mitigating attacks such as Membership Inference and Gradient Inversion.

Despite its strong theoretical guarantees, Differential Privacy introduces an inherent trade-off between privacy and predictive performance. Increasing the amount of injected noise generally improves privacy protection but may simultaneously reduce model accuracy and slow convergence. Consequently, selecting an appropriate privacy budget remains one of the principal challenges when deploying Differential Privacy in practical federated learning systems.

Recent research has focused on developing adaptive privacy mechanisms, personalized privacy budgets and client-specific noise allocation strategies to improve this trade-off while maintaining rigorous privacy guarantees [44] [49]. As a result, Differential Privacy remains one of the most actively investigated privacy-preserving mechanisms for federated learning, particularly in sensitive domains such as healthcare.

Homomorphic Encryption

Homomorphic Encryption (HE) is a cryptographic technique that enables computations to be performed directly on encrypted data without requiring prior decryption. Unlike conventional encryption schemes, where data must be decrypted before any processing can occur, Homomorphic Encryption preserves data confidentiality throughout the computation process. As a result, untrusted entities can execute mathematical operations over encrypted information while remaining unable to access the underlying plaintext, making HE particularly attractive for privacy-sensitive applications such as healthcare and federated learning [1] [31].

The concept of fully homomorphic encryption was first introduced by Gentry [20], who demonstrated that arbitrary computations could be performed on encrypted data while preserving correctness after decryption. Although this breakthrough established the theoretical feasibility of privacy-preserving computation, the original scheme incurred extremely high computational costs, limiting its practical adoption.

Subsequent research has focused on improving the efficiency of homomorphic encryption schemes. Among the most widely adopted approaches is the Cheon–Kim–Kim–Song (CKKS) scheme proposed by Cheon et al. [11]. Unlike earlier exact arithmetic schemes, CKKS supports approximate arithmetic over real-valued numbers, making it particularly suitable for machine learning applications where small numerical approximations have negligible impact on predictive performance. Its ability to efficiently process floating-point operations has made CKKS one of the dominant encryption schemes for privacy-preserving federated learning.

Within federated learning, Homomorphic Encryption is primarily employed to protect model parameters or gradients exchanged between clients and the aggregation server. Before transmission, each client encrypts its local model updates using a public encryption key. The server subsequently performs aggregation directly on the encrypted parameters without accessing their plaintext values. Once aggregation is completed, the encrypted global model can be decrypted only by authorized participants possessing the corresponding secret key. This approach prevents curious servers and external adversaries from inspecting the transmitted model updates while preserving the correctness of the aggregation process [42] [33].

Despite these advantages, Homomorphic Encryption introduces non-negligible computational and communication overhead. Encryption, ciphertext expansion, and encrypted arithmetic substantially increase memory consumption and execution time compared with conventional federated learning. Consequently, recent research has focused on improving the efficiency of homomorphic encryption through optimized cryptographic libraries, ciphertext compression techniques, partial parameter encryption, and hybrid privacy-preserving frameworks that combine Homomorphic Encryption with complementary mechanisms such as Differential Privacy and secure aggregation [60] [33].

Among the various encryption schemes currently available, CKKS has emerged as the preferred choice for privacy-preserving federated learning involving deep neural networks because it provides an effective compromise between computational efficiency and encryption capability. Ongoing research continues to investigate methods for reducing its computational overhead while preserving the strong confidentiality guarantees required

for sensitive applications such as collaborative healthcare analytics.

2.6.5 Robust Aggregation Strategies

The effectiveness of federated learning relies heavily on the aggregation strategy employed by the central server to combine the model updates received from participating clients. The aggregation algorithm not only determines the convergence behaviour and predictive performance of the global model but also influences the system's resilience against malicious or faulty client updates. Consequently, designing aggregation methods that are both efficient and robust has become a major research challenge in federated learning [30] [38].

The most widely adopted aggregation algorithm is Federated Averaging (FedAvg), introduced by McMahan et al. [36]. FedAvg computes the weighted average of the model parameters received from the participating clients, where each contribution is weighted according to the size of the corresponding local dataset. Owing to its simplicity, low computational complexity, and excellent convergence properties, FedAvg has become the de facto baseline for federated learning systems.

Despite its widespread adoption, FedAvg assumes that all participating clients behave honestly and that their model updates are reliable. In adversarial environments, however, malicious clients may deliberately manipulate their local updates to influence the aggregation process, inject backdoors, or degrade the performance of the global model. Since FedAvg assigns weights solely according to dataset size, it lacks mechanisms for identifying or mitigating abnormal client behaviour, making it particularly vulnerable to poisoning attacks [3].

To address these limitations, numerous Byzantine-robust aggregation algorithms have been proposed. Early approaches include Krum [7], which selects the client update that is closest to the majority of the remaining updates, and coordinate-wise aggregation methods such as Median [56] and Trimmed Mean [56], which reduce the influence of outliers by computing robust statistics independently for each model parameter. Bulyan [14] further combines robust selection and aggregation strategies to improve resistance against sophisticated Byzantine attacks.

Although these methods improve robustness against malicious updates, they generally treat all model parameters equally and therefore neglect the structural characteristics of deep neural networks. In convolutional neural networks, however, convolutional kernels capture hierarchical spatial features that are particularly important for image classification tasks. Consequently, analysing the similarity of convolutional kernels rather than individual parameters may provide a more reliable indication of whether a client update is consistent with the majority of honest participants.

Motivated by this observation, Zhu et al. [62] proposed the Convolutional Kernel Angle-Based Defense Aggregation (CKADA) algorithm. Instead of relying solely on parameter magnitudes or coordinate-wise statistics, CKADA evaluates the angular similarity between convolutional kernel updates received from different clients. Client updates exhibiting abnormal kernel orientations are identified as potentially malicious and excluded from the aggregation process before constructing the global model. By exploiting the structural properties of convolutional neural networks, CKADA demonstrates improved robustness against poisoning and backdoor attacks while maintaining competitive

predictive performance on benign data.

The continuous evolution of robust aggregation algorithms reflects the growing need to defend federated learning systems against increasingly sophisticated adversarial attacks. Current research is therefore moving towards hybrid privacy-preserving frameworks that combine robust aggregation with complementary techniques such as Differential Privacy and Homomorphic Encryption. These integrated approaches seek to simultaneously improve privacy, robustness, and communication security, representing one of the most active research directions in secure federated learning.

The literature reviewed in this section demonstrates that, although federated learning significantly improves data privacy by eliminating the need for centralized data collection, it remains vulnerable to a wide range of privacy and security threats. Membership Inference, Gradient Inversion, and Backdoor attacks illustrate how sensitive information can be inferred or how the collaborative learning process can be manipulated by malicious participants. At the same time, defence mechanisms such as Differential Privacy, Homomorphic Encryption, and robust aggregation strategies provide complementary solutions that address different aspects of these challenges, each introducing distinct trade-offs between privacy protection, robustness, computational overhead, and predictive performance.

Building upon these findings, the next chapter presents the design and implementation of the proposed privacy-preserving federated learning framework, which integrates these complementary mechanisms into a unified architecture and experimentally evaluates their effectiveness under representative adversarial scenarios.

Chapter 3

Methodology and Experimental Design

3.1 Research Methodology

This research adopts an experimental methodology to design, implement and evaluate a privacy-preserving federated learning framework for medical image classification. The proposed framework integrates Differential Privacy (DP), Homomorphic Encryption (HE), and robust aggregation techniques to improve the privacy and security of collaborative model training while maintaining acceptable predictive performance.

The study follows a systematic workflow comprising five main phases, illustrated in Figure 3.1. First, an extensive review of the existing literature was conducted to identify the principal privacy and security challenges affecting federated learning in healthcare, as well as the most widely adopted countermeasures. The findings of this review informed the design of the proposed framework and the selection of the privacy-preserving mechanisms investigated in this work.

The second phase consisted of designing the federated learning system. A client–server architecture was implemented using the Flower framework [6], with the BloodMNIST dataset [54] distributed across multiple simulated healthcare institutions. Different data partitioning strategies were considered to reproduce both homogeneous and heterogeneous data distributions typically encountered in real-world federated learning environments.

In the third phase, the proposed privacy-preserving framework was developed by integrating Differential Privacy, Homomorphic Encryption and robust aggregation into the federated training pipeline. Each component was implemented in a modular manner, allowing individual mechanisms to be enabled or disabled independently throughout the experimental campaign.

The fourth phase focused on evaluating the framework under different experimental conditions. Several federated learning configurations were investigated by varying the aggregation strategy, the privacy-preserving mechanisms and the client data distribution. A centralized learning model was also trained to provide a performance baseline for comparison.

Finally, the framework was assessed from three complementary perspectives: predictive performance, computational efficiency and security. Utility was evaluated using standard classification metrics, computational performance was analysed through execution time, communication overhead and computational cost, while security was assessed by performing Membership Inference, Frequency-Injection Backdoor and Gradient Inversion attacks against the trained models.

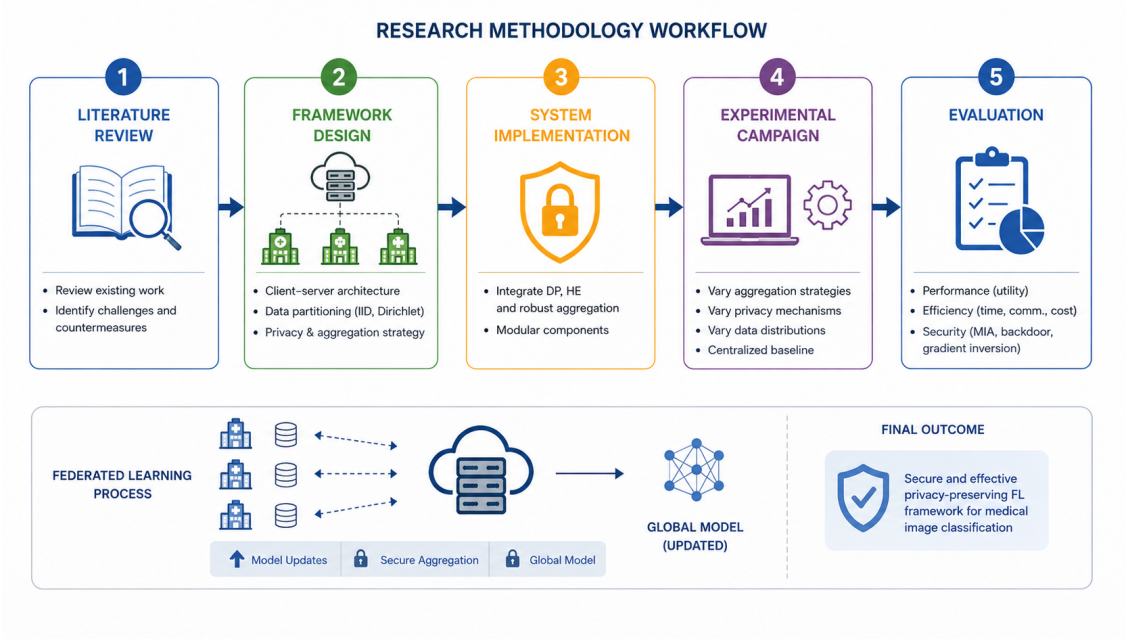


Figure 3.1: Research workflow illustrating the methodology adopted for the design, implementation, and evaluation of the proposed privacy-preserving federated learning framework.

3.2 System Architecture

This section describes the overall architecture of the proposed federated learning framework. The system follows a centralized client–server topology in which multiple healthcare institutions collaboratively train a shared deep learning model without exchanging their local datasets. Instead of transmitting raw medical data, only model parameters or model updates are communicated between the participating clients and the central server, thereby preserving data locality throughout the training process.

The proposed framework was implemented using the Flower federated learning framework, which provides the communication infrastructure and orchestration mechanisms required to coordinate distributed model training. The architecture was designed to support different aggregation strategies and privacy-preserving mechanisms in a modular manner, allowing the experimental evaluation of multiple federated learning configurations.

Figure 3.2 illustrates the overall architecture of the proposed framework.

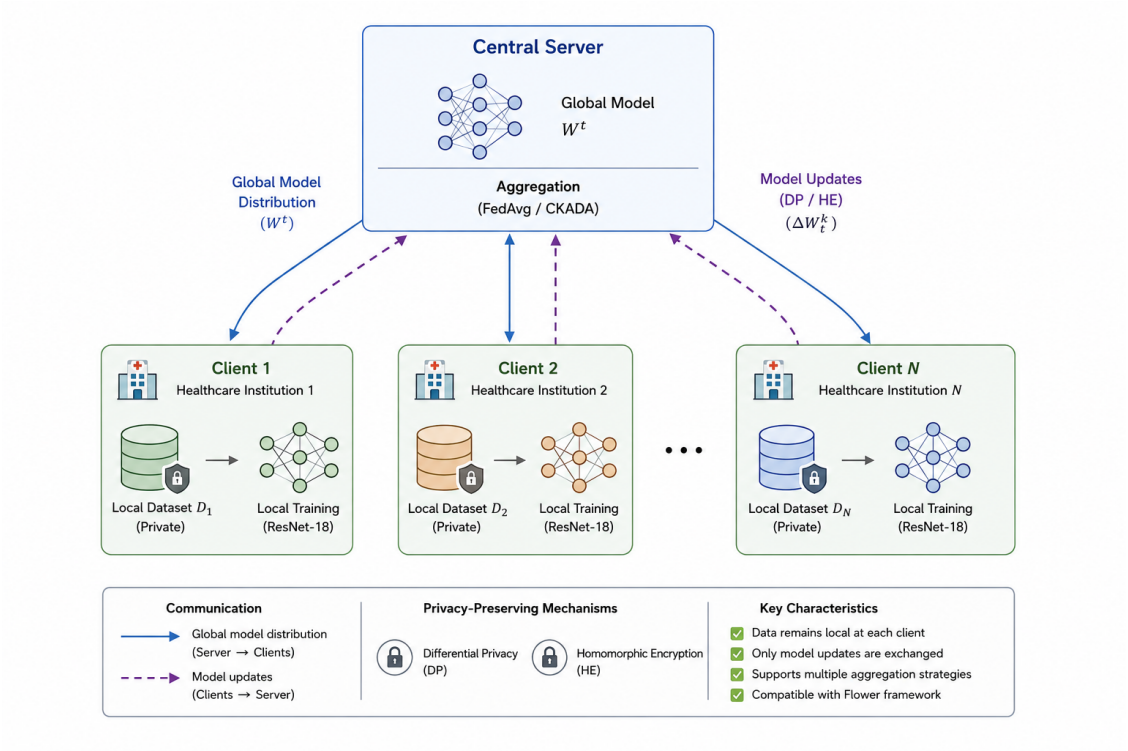


Figure 3.2: Overall architecture of the proposed federated learning framework, illustrating the interaction between the central server and multiple healthcare clients during collaborative model training.

3.2.1 Overall Architecture

The federated learning environment consists of one central server and multiple participating clients, each representing an independent healthcare institution. Every client stores its own local dataset and performs model training independently using local computational resources. The server coordinates the federated learning process by distributing the global model at the beginning of each communication round, collecting the corresponding client updates after local training, aggregating the received information, and generating the updated global model for the next training round.

Unlike centralized machine learning, patient data never leaves the local institutions. This design significantly reduces the risks associated with centralized data storage while enabling collaborative model training across geographically distributed organizations.

The architecture adopted in this work follows the standard cross-silo federated learning paradigm, where a limited number of reliable institutions participate throughout the entire training process. In the experimental evaluation presented in this thesis, three clients collaborate with a single central server.

3.2.2 Flower-Based Federated Learning Framework

The proposed system was implemented using the Flower framework, an open-source platform for federated learning research and deployment. Flower provides the communication infrastructure required to coordinate distributed model training while remaining independent of the underlying deep learning framework.

Within the proposed implementation, the server is responsible for initializing the global model, selecting the participating clients, distributing the current model parameters, collecting client updates and invoking the selected aggregation strategy. The clients receive the global model, perform local optimization using their private datasets, and return the resulting model updates to the server at the end of each communication round.

One advantage of Flower is its modular strategy interface, which allows different aggregation algorithms to be integrated without modifying the client implementation. This capability enables the experimental comparison of multiple aggregation strategies, including FedAvg, CKADA and the proposed hybrid aggregation mechanism described later in this chapter.

3.2.3 Communication Workflow

Each communication round follows the same sequence of operations. Initially, the server distributes the current global model to all selected clients. Every client subsequently performs local training using its private dataset and generates a local model update. Once local optimization is completed, the updates are transmitted back to the server, where the selected aggregation strategy combines the received information to produce the next global model.

This communication cycle is repeated until the predefined stopping criterion is satisfied. Throughout the entire training process, only model information is exchanged between clients and the server, while the original training data remain exclusively within the corresponding healthcare institutions.

The complete federated training procedure is presented in Algorithm 1, whereas the implementation details of the aggregation strategies and privacy-preserving mechanisms are described in Sections 3.5 and 3.6, respectively.

3.3 Dataset Description

The experimental evaluation conducted in this work is based on BloodMNIST, a publicly available medical image-classification dataset included in the MedMNIST collection [53] [54]. BloodMNIST contains microscopic images of individual peripheral blood cells and defines an eight-class supervised classification problem.

The dataset was selected because it provides a suitable balance between medical relevance, experimental reproducibility, and computational feasibility. Its relatively compact image size makes it possible to execute a large number of federated learning experiments, including configurations involving Differential Privacy, Homomorphic Encryption, robust aggregation, and privacy or security attacks. At the same time, its multi-class structure

and naturally unbalanced class distribution provide a sufficiently challenging learning problem.

This section first introduces the characteristics of BloodMNIST, then describes the preprocessing operations applied before training, and finally presents the IID and non-IID partitioning strategies used to distribute the data among the three federated clients.

3.3.1 Dataset Selection

The experiments presented in this thesis were conducted using the BloodMNIST dataset from the MedMNIST v2 benchmark collection [54]. MedMNIST is a standardized benchmark designed for biomedical image classification tasks and provides lightweight datasets suitable for evaluating machine learning and federated learning algorithms under reproducible conditions.

BloodMNIST is a multi-class image classification dataset composed of microscopic images of peripheral blood cells collected from healthy individuals. Representative samples from the eight cell categories are shown in Figure 3.3. As summarized in Table 3.1, the dataset contains 17,092 RGB images with a spatial resolution of 28×28 pixels, divided into official training, validation, and test splits comprising eight clinically relevant cell categories.

The dataset was selected because it represents a realistic healthcare image classification task while remaining computationally efficient for large-scale federated learning experiments involving multiple aggregation strategies, privacy-preserving mechanisms, and security evaluations.

Table 3.1: Official BloodMNIST dataset splits.

Dataset split	**Number of images**	**Percentage**
Training	11,959	69.97%
Validation	1,712	10.02%
Test	3,421	20.01%
Total	17,092	100.00%

Figure 3.3 presents representative examples from the eight BloodMNIST classes. Although the images have a relatively low spatial resolution (28×28 pixels), they preserve the morphological characteristics required to distinguish between the different blood cell types. These representative samples illustrate the diversity of the dataset and highlight the complexity of the multi-class classification task.

3.3.2 Data Preprocessing

Before training, all images were transformed into PyTorch tensors [39] and normalized using the standard preprocessing pipeline adopted for the BloodMNIST dataset. No data augmentation techniques were applied in order to ensure a consistent comparison across all experimental configurations.

The same preprocessing pipeline was employed for both centralized and federated learning experiments, guaranteeing that differences in performance originated solely from

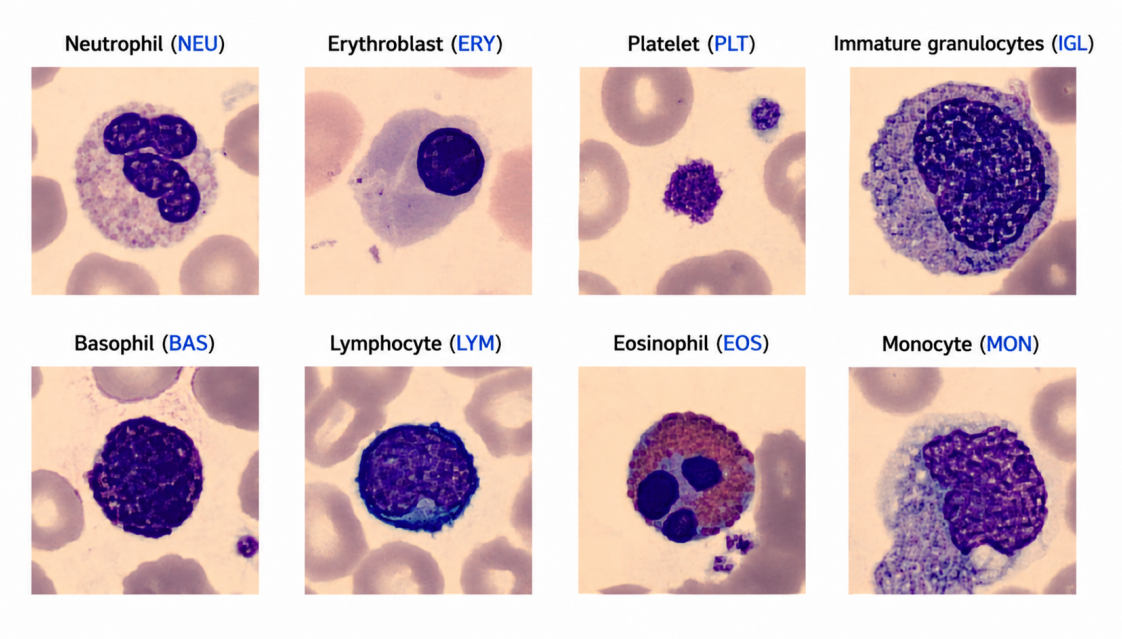


Figure 3.3: Representative samples from the eight classes of the BloodMNIST dataset. The images correspond to the cell categories included in the MedMNIST benchmark and illustrate the visual variability of the classification task.

the learning strategy and privacy-preserving mechanisms rather than from variations in data preparation.

3.3.3 Data Partitioning

To simulate distributed healthcare institutions, the official BloodMNIST training set was partitioned among three federated clients. Each client retained its local data throughout the training process, while only model updates were exchanged with the central server.

Three partitioning strategies were considered to evaluate the influence of data heterogeneity on federated learning performance:

- **IID Partitioning.** The training samples were randomly distributed among the participating clients while preserving approximately the same class distribution in every local dataset. This configuration represents the ideal federated learning scenario in which each institution observes similar data distributions.
- **Dirichlet Partitioning ($\alpha = 0.5$).** Client datasets were generated using a Dirichlet distribution with concentration parameter $\alpha = 0.5$, resulting in a moderate level of statistical heterogeneity between institutions.
- **Dirichlet Partitioning ($\alpha = 0.1$).** A smaller concentration parameter produced a highly heterogeneous distribution in which certain classes became concentrated within a limited number of clients. This configuration represents a more challenging

and realistic federated learning scenario commonly encountered in clinical environments.

In all federated learning experiments, the official BloodMNIST training split containing 11,959 images was partitioned among three participating clients. As summarized in Table 3.2, under the IID configuration each client received approximately one third of the training data, corresponding to about 3,986 images. Subsequently, each local partition was divided into 80% for local training and 20% for local validation, resulting in approximately 3,189 training samples and 797 validation samples per client. The same local train/validation split was applied to all partitioning strategies to ensure consistent experimental conditions.

Table 3.2: Dataset partitioning for each federated client under the IID configuration.

Description	**Samples**
Total BloodMNIST training images	11,959
Number of federated clients	3
Images per client partition	$\approx$ 3,986
Local training set (80%)	$\approx$ 3,189
Local validation set (20%)	$\approx$ 797

The official BloodMNIST test set remained unchanged throughout all experiments and was used exclusively for evaluating the global model after each communication round.

3.4 Model Architecture

This section describes the convolutional neural network employed for the image classification task and the training configuration adopted throughout the experimental evaluation. The same model architecture and optimization settings were used for all experiments to ensure that differences in performance could be attributed exclusively to the federated learning configuration and privacy-preserving mechanisms rather than changes to the underlying classifier.

3.4.1 Neural Network Architecture

The image classification model adopted in this work is based on the ResNet-18 architecture, a deep convolutional neural network introduced by He et al. [25]. ResNet-18 was selected because it provides a favourable balance between predictive performance, computational efficiency and model complexity, making it well suited for federated learning environments where communication cost and local computational resources are important considerations.

A distinguishing characteristic of ResNet architectures is the use of residual learning. Instead of learning a direct mapping between the input and output of each block, the network learns a residual function while shortcut connections allow the original input

to bypass intermediate convolutional layers. These identity mappings facilitate gradient propagation during backpropagation, mitigate the vanishing gradient problem and enable deeper networks to be trained more effectively.

The architecture of the modified ResNet-18 model adopted in this work is summarized in Table 3.3. In particular, the input stem was adapted to the small spatial resolution of BloodMNIST images by using a 3×3 convolution followed by Group Normalization and ReLU activation. This avoids the aggressive early downsampling typically used in the standard ResNet-18 architecture and helps preserve spatial information in the input images.

The implemented network consists of an initial convolutional layer followed by four residual stages, each composed of two Basic Residual Blocks. The spatial resolution is progressively reduced while the number of feature channels increases, allowing the network to learn hierarchical image representations. After the final residual stage, global average pooling aggregates the extracted feature maps into a fixed-length feature vector, which is passed to a fully connected layer producing the final class probabilities for the eight BloodMNIST categories.

Compared with larger residual networks, ResNet-18 offers a significantly lower computational cost while maintaining competitive classification performance. This characteristic is particularly advantageous in federated learning, where every participating client performs local model optimization independently during each communication round.

Table 3.3: Modified ResNet-18 architecture adopted for BloodMNIST classification.

Stage	**Output channels**	**Structure**
Input	3	BloodMNIST RGB image
Input stem	64	3×3 convolution, GN, ReLU
Residual stage 1	64	2 basic residual blocks
Residual stage 2	128	2 basic residual blocks, downsampling
Residual stage 3	256	2 basic residual blocks, downsampling
Residual stage 4	512	2 basic residual blocks, downsampling
Pooling	512	Adaptive average pooling
Classification layer	8	Fully connected layer

3.4.2 Training Configuration

The same training configuration was employed throughout all experiments to ensure a fair comparison between the evaluated federated learning scenarios. Model optimization was performed locally on each participating client using stochastic gradient descent (SGD), while the global model was obtained through the aggregation strategy selected for each experiment.

Table 3.4 summarises the principal hyperparameters adopted during local training.

The selected hyperparameters were chosen to provide a stable and consistent training configuration across all experimental scenarios. A learning rate of 0.01 was selected to ensure stable optimization while avoiding oscillations during local training. Each

Table 3.4: Training configuration adopted for the classification model.

Parameter	**Value**
Model architecture	ResNet-18
Number of output classes	8
Optimizer	Stochastic Gradient Descent
Learning rate	0.01
Batch size	32
Local epochs	10
Loss function	Cross-entropy loss
Random seed	327834

client performed 10 local epochs per communication round, providing sufficient local model improvement without introducing excessive client drift. A batch size of 32 offered a suitable compromise between computational efficiency and gradient stability, while the cross-entropy loss function was employed as it is the standard objective for multi-class image classification tasks. Finally, a fixed random seed was used throughout all experiments to guarantee reproducibility and enable fair comparisons between different federated learning configurations.

The hardware platform used for all experiments is summarized in Table 3.5. Unless otherwise stated, all experiments were executed on the same system to ensure fair and reproducible comparisons.

Table 3.5: Experimental hardware environment.

Component	**Specification**
Operating System	macOS (Build 25F84)
CPU	Apple M3
Memory	16 GB unified memory
GPU	Apple M3 Integrated GPU
Acceleration Backend	PyTorch MPS (Metal Performance Shaders)

3.5 Federated Learning Configuration

This section describes the federated learning configuration adopted in this work, including the software framework used to orchestrate distributed training, the aggregation strategies implemented at the server, and the overall training procedure. The objective is to establish a common federated learning pipeline that serves as the foundation for the privacy-preserving mechanisms introduced in the following section.

3.5.1 Flower Framework

The proposed federated learning framework was implemented using Flower, an open-source framework designed to simplify the development and evaluation of distributed machine learning systems. Flower provides the communication infrastructure between the central server and participating clients while remaining independent of the underlying deep learning library.

Within the proposed implementation, the server is responsible for initializing the global model, coordinating each communication round, distributing the current global model parameters, collecting the local model updates, and invoking the selected aggregation strategy to construct the updated global model. In all experiments, every available client participates in every communication round; therefore, no client sampling or partial participation strategy is employed. This design ensures that each global model update incorporates information from the complete distributed dataset, although the underlying data remain locally stored at the participating institutions.

Each client performs local model optimization using its private dataset. After receiving the current global model, the client executes the predefined number of local training epochs and returns the updated model parameters to the server. Since all clients participate in every communication round, each aggregation step combines model updates from every local dataset. Throughout the entire training process, only model information is exchanged, while the local training data remain exclusively on the corresponding client.

The modular architecture provided by Flower enables different aggregation strategies to be integrated without modifying the client-side training procedure. This flexibility allows the proposed framework to evaluate multiple aggregation mechanisms under identical experimental conditions.

3.5.2 Aggregation Strategies

The aggregation strategy executed by the server determines how the local client models are combined to produce the updated global model after each communication round. As summarized in Table 3.6, three aggregation strategies were implemented and evaluated in this work.

The first strategy is Federated Averaging (FedAvg), proposed by McMahan et al. [36]. FedAvg computes the weighted average of the model parameters received from all participating clients according to the size of their local datasets. Owing to its simplicity, communication efficiency, and widespread adoption, it serves as the baseline aggregation algorithm throughout this work.

The second strategy is CKADA [62], a robust aggregation algorithm designed to improve resilience against malicious client updates. Unlike conventional parameter averaging, CKADA evaluates the consistency of convolutional kernel updates before aggregation, reducing the influence of abnormal or potentially poisoned model updates contributed by compromised clients.

To support the proposed privacy-preserving framework presented in Section 3.6, a hybrid aggregation mechanism was also implemented. In this approach, CKADA is applied to the subset of model parameters protected by Differential Privacy, while the remaining encrypted parameters are aggregated using FedAvg. This design enables robust aggregation and encrypted parameter aggregation to coexist within the same federated learning framework.

3.5.3 Federated Training Procedure

Federated training proceeds iteratively through a sequence of communication rounds coordinated by the central server. At the beginning of each round, the server distributes the current global model to all participating clients. Each client independently performs local optimization using its private dataset and returns the resulting model update to the server.

After collecting the updates from all clients, the server applies the selected aggregation strategy to construct the updated global model. The resulting model is then redistributed during the next communication round, and the process continues until the predefined stopping criterion is satisfied.

Table 3.6: Aggregation strategies implemented in the proposed framework.

Strategy	**Aggregation principle**	**Primary objective**
FedAvg	Weighted averaging of local model parameters	Baseline federated aggregation with low computational overhead
CKADA	Kernel-angle based aggregation of convolutional updates	Improve robustness against malicious or abnormal client updates
Hybrid (CKADA + FedAvg)	CKADA on selected layers, FedAvg on remaining layers	Support simultaneous robust aggregation and encrypted parameter aggregation

Algorithm 1 summarizes the complete federated training procedure adopted throughout the experimental evaluation. Unlike many federated learning implementations that perform client sampling at each communication round, the proposed framework adopts a full-participation strategy in which all clients contribute to every aggregation step.

Algorithm 1 Flower-based Federated Training Procedure

Require: Initial global model W^0, clients $\{C_1, \ldots, C_K\}$, number of communication rounds R

for $t = 0, \ldots, R-1$ **do**
  Server distributes global model W^t to all clients
  for all clients C_k **in parallel do**
    Load W^t into the local model
    Train using local dataset D_k
    Generate updated local model parameters
    Return local model update to the server
  end for
  Aggregate client updates using the selected aggregation strategy
  Generate updated global model W^{t+1}
end for
return Final global model

3.6 Proposed Privacy-Preserving Federated Learning Framework

The federated learning configuration described in the previous section provides the foundation for collaborative model training. This section presents the proposed privacy-preserving framework developed in this thesis, which extends the baseline federated learning pipeline through the integration of Differential Privacy (DP), Homomorphic Encryption (HE), and robust aggregation.

The proposed framework was designed with three primary objectives. First, it aims to reduce the risk of information leakage from transmitted model updates by perturbing sensitive gradients using Differential Privacy. Second, it protects the confidentiality of selected model parameters during transmission and aggregation through Homomorphic Encryption. Finally, it maintains robustness against malicious client updates by combining robust aggregation with encrypted model aggregation within a unified training pipeline.

Unlike conventional federated learning systems, where all model parameters are aggregated identically, the proposed framework adopts a hybrid aggregation strategy. Parameters protected by Differential Privacy are aggregated using CKADA, while encrypted parameters are aggregated using Federated Averaging. This design enables both privacy-preserving mechanisms to coexist without violating the computational requirements of either aggregation strategy.

3.6.1 Design Objectives

The proposed framework was developed to balance three competing requirements commonly encountered in privacy-preserving federated learning: predictive performance, privacy protection, and computational efficiency.

From a utility perspective, the framework should maintain classification performance comparable to conventional federated learning despite the introduction of privacy-preserving mechanisms. From a security perspective, it should reduce the effectiveness of privacy attacks while increasing the difficulty of reconstructing sensitive patient information from exchanged model updates. Finally, the framework should remain computationally practical by limiting the communication and computational overhead introduced during collaborative training.

To satisfy these objectives, the proposed framework adopts a modular architecture in which each privacy-preserving component can be independently enabled or disabled. This modular design facilitates a systematic experimental evaluation of the individual and combined contributions of Differential Privacy and Homomorphic Encryption.

3.6.2 Differential Privacy Integration

Differential Privacy (DP) was integrated into the client-side training procedure using the Opacus library [58]. The implementation adopts Differentially Private Stochastic Gradient Descent (DP-SGD), in which privacy protection is enforced during local optimization before model updates are transmitted to the central server.

Before enabling Differential Privacy, the ResNet-18 architecture is validated using the Opacus `ModuleValidator`. Layers that are incompatible with per-sample gradient computation, particularly Batch Normalization layers, are automatically replaced with compatible alternatives such as Group Normalization. This modification is required because Batch Normalization computes statistics across multiple samples within the same mini-batch, thereby introducing dependencies between individual records that are incompatible with standard sample-level Differential Privacy assumptions.

During local training, Opacus computes gradients separately for each training sample. Let g_i denote the gradient generated by sample i. Each per-sample gradient is clipped according to the maximum norm C:

$$\bar{g}_i = g_i \cdot \min\left(1, \frac{C}{\|g_i\|_2}\right). \tag{3.1}$$

As defined in Equation 3.1, the clipping operation bounds the contribution of each individual training sample and therefore limits the sensitivity of the optimization procedure. After clipping, the gradients are averaged and Gaussian noise is added before the optimizer updates the local model parameters:

$$\tilde{g} = \frac{1}{B}\left(\sum_{i=1}^{B} \bar{g}_i + \mathcal{N}\left(0, \sigma^2 C^2 I\right)\right), \tag{3.2}$$

Equation 3.2 defines the perturbed gradient used during local optimization, where B denotes the effective batch size, C is the clipping threshold, and σ is the noise multiplier

controlling the amount of injected Gaussian noise. The resulting perturbed gradient $\tilde{g}$ is then used by the optimizer to update the local model parameters.

To reduce the memory overhead associated with per-sample gradient computation, the implementation enables Ghost Clipping [9]. This technique computes the norms required for clipping without explicitly storing the complete per-sample gradient tensors, thereby making DP-SGD more practical for deep convolutional networks such as ResNet-18.

The implementation also supports gradient accumulation to obtain a larger effective batch size without exceeding the available memory. Local mini-batches are processed using unreduced losses, while intermediate accumulation steps are performed through the Opacus `virtual_step` mechanism. Once the predefined number of accumulation steps is reached, the optimizer applies the clipped and noisy gradient update. This approach improves the signal-to-noise ratio while preserving the privacy guarantees enforced by the Privacy Engine.

The privacy level is controlled through the target privacy budget ε and the parameter δ. Lower values of ε correspond to stronger privacy guarantees and generally require a larger amount of injected noise, whereas higher values provide weaker privacy protection but typically preserve model utility more effectively.

At the end of each local training round, the effective privacy expenditure is estimated using the Opacus privacy accountant. The resulting value of ε, computed for the selected δ, is returned to the Flower server through the client metrics. This allows the privacy budget consumed during training to be monitored across communication rounds.

Within the proposed framework, Differential Privacy can be selectively enabled or disabled. This modular design enables its effect on predictive performance, convergence behaviour, computational overhead, Membership Inference resistance, and Gradient Inversion resistance to be evaluated independently.

3.6.3 Homomorphic Encryption Integration

Homomorphic Encryption (HE) was incorporated into the proposed framework using the CKKS approximate arithmetic scheme implemented through the TenSEAL library [5]. Rather than encrypting the complete model, the proposed framework adopts a *selective encryption* strategy, allowing only the most privacy-sensitive model layers to be protected while reducing the computational and communication overhead associated with fully homomorphic aggregation.

Before the federated training process begins, a CKKS encryption context is generated using the parameters reported in Table 3.7. The public encryption context is distributed to all participating clients, while the corresponding secret key is retained locally and used exclusively for decryption operations.

To identify the layers requiring protection, the framework performs a sensitivity evaluation phase during the first communication round. Each client executes one epoch of local training and computes a sensitivity score for every model layer as the mean absolute difference between the initial and locally updated parameters. These scalar sensitivity values are encrypted using CKKS and transmitted to the server, where they are aggregated directly in the encrypted domain. After decryption of the aggregated sensitivities, the server ranks the model layers according to their average sensitivity and generates a

global encryption mask.

The proportion of encrypted layers is controlled by the configurable parameter `he_ratio`, which specifies the percentage of the most sensitive layers to be protected. In addition, the first convolutional layer and the final classification layer are always included in the encryption mask, irrespective of their measured sensitivity, since these layers are generally considered the most susceptible to feature and label leakage.

During the subsequent federated learning rounds, clients encrypt only the layers selected by the global mask before transmitting their model updates to the server. Encrypted parameters are serialized together with their tensor metadata, while the remaining layers are transmitted in plaintext. Upon receiving the mixed updates, the server identifies the encrypted parameters and performs the weighted Federated Averaging operations directly on the ciphertexts using the homomorphic addition and scalar multiplication operations supported by the CKKS scheme. Consequently, the server never accesses the plaintext values of the protected parameters during model aggregation.

After aggregation, the encrypted global parameters are returned to the clients, where they are decrypted using the corresponding secret key and reconstructed into their original tensor representation before the next communication round begins. This sensitivity-based selective encryption strategy enables the proposed framework to achieve a balance between privacy protection and computational efficiency while preserving compatibility with the federated learning workflow.

Table 3.7: CKKS encryption parameters adopted in the proposed framework.

Parameter	**Value**
Encryption library	TenSEAL
Encryption scheme	CKKS
Polynomial modulus degree	8192
Coefficient modulus sizes	{60,40,40,60} bits
Global scale	2^{40}
Selective encryption	Sensitivity-based layer selection
Encryption ratio	he_ratio
Protected layers	Top sensitive layers + first and last layer

3.6.4 Secure Training Workflow

Figure 3.4 illustrates the complete training pipeline adopted by the proposed framework. At the beginning of each communication round, the server distributes the current global model to all participating clients. Each client performs local optimization using its private dataset and subsequently applies the configured privacy-preserving mechanisms.

When Differential Privacy is enabled, gradients are clipped and perturbed before the local model update is generated. When Homomorphic Encryption is enabled, the selected model parameters are encrypted before transmission to the server. After receiving the client updates, the server performs aggregation using the hybrid strategy described in

Section 3.5. Finally, the updated global model is redistributed to all clients, initiating the next communication round.

This modular workflow enables each privacy-preserving mechanism to be evaluated independently or in combination, providing a flexible experimental platform for analysing the trade-offs between predictive performance, computational overhead, and security.

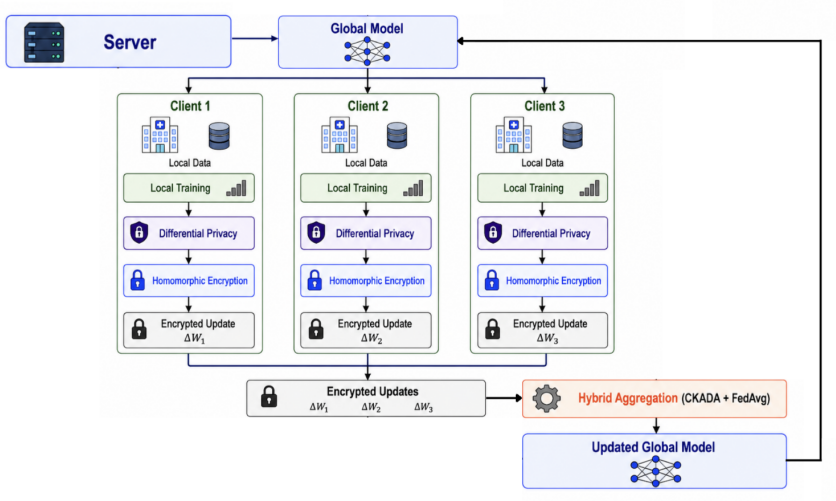


Figure 3.4: Secure federated training workflow of the proposed framework. Each client performs local training on private data, computes model updates, applies Differential Privacy and Homomorphic Encryption, and transmits the protected updates to the server, where hybrid aggregation (CKADA + FedAvg) is performed to construct the updated global model.

3.7 Security Threat Model and Attack Implementation

The objective of this section is to describe the adversarial scenarios considered to assess the security of the proposed privacy-preserving federated learning framework. As discussed in Chapter 2, federated learning systems are susceptible to various privacy and security threats that may compromise either the confidentiality of the training data or the integrity of the collaboratively trained model. To evaluate the effectiveness of the proposed framework against representative attacks, three attack classes were implemented and integrated into the experimental pipeline.

The selected attacks investigate complementary aspects of federated learning security. The Membership Inference Attack evaluates whether an adversary can determine if a specific data sample participated in the training process, thereby assessing the privacy leakage of the global model. The Frequency-Injection Backdoor Attack examines the robustness of the aggregation process against malicious clients attempting to manipulate the model behaviour through poisoned local updates. Finally, the Gradient Inversion Attack evaluates whether sensitive training samples can be reconstructed from the information exchanged during collaborative training.

Rather than proposing new attack methodologies, this work integrates state-of-the-art attacks from the literature into the developed Flower-based federated learning framework. This enables a systematic evaluation of the proposed privacy-preserving mechanisms under realistic adversarial conditions while ensuring that the obtained results remain directly comparable with previous studies.

The remainder of this section first defines the threat model considered throughout the experimental campaign and subsequently describes the implementation of each attack within the proposed framework.

3.7.1 Threat Model

The security evaluation presented in this thesis assumes a cross-silo federated learning environment composed of a central server and three participating clients representing independent healthcare institutions. Each client possesses a private local dataset that remains on-premises throughout the entire training process, while model parameters or updates are exchanged with the central server according to the selected aggregation strategy.

The central server is assumed to follow the prescribed federated learning protocol correctly. Consequently, the evaluation focuses on threats originating from malicious or curious clients, which represent a realistic security concern in collaborative learning environments involving multiple independent organizations.

Three adversarial capabilities are considered:

- **Curious adversary.** The attacker attempts to infer sensitive information about the local training data from the exchanged model updates or the trained global model without modifying the federated learning protocol. This capability is evaluated through the Membership Inference Attack and the Gradient Inversion Attack.

- **Malicious participant.** One participating client intentionally manipulates its local training data to introduce a backdoor into the global model while remaining indistinguishable from honest participants. This capability is evaluated using the Frequency-Injection Backdoor Attack.

- **Knowledge assumptions.** The attacker is assumed to possess knowledge of the model architecture and the federated learning protocol but has no direct access to the private datasets stored by other clients.

This threat model reflects the assumptions commonly adopted in the federated learning literature and provides a realistic basis for evaluating the effectiveness of the proposed privacy-preserving mechanisms.

3.7.2 Membership Inference Attack

Membership Inference Attacks (MIAs) aim to determine whether a specific data sample was used during the training of a machine learning model. As discussed in Chapter 2,

successful membership inference may reveal sensitive information regarding the participation of individual patients in the collaborative training process, representing a significant privacy risk for healthcare applications.

To evaluate the resistance of the proposed framework against this threat, the Quantile Membership Inference Attack proposed by Watson et al. [47] was implemented. This attack was selected because it provides a scalable and model-agnostic evaluation methodology that is suitable for deep learning models trained in federated environments.

Integration into the Proposed Framework

The attack was executed after the federated training process had converged. The final global model was considered the target model, while additional datasets were prepared to reproduce the attack pipeline proposed by Watson et al.

Following the original methodology, four datasets were employed:

- the final global federated model acting as the target model;
- the official BloodMNIST test split, used as the public dataset for quantile regression training;
- the training partition of Client 0, representing the member dataset;
- the testing partitions of Clients 1 and 2, used respectively for calibration and non-member evaluation.

Using these datasets, quantile regression models were trained to estimate the expected distribution of prediction scores for non-member samples. The resulting quantile estimates were subsequently used to infer the membership status of unseen samples by comparing their prediction scores with the estimated decision thresholds.

Experimental Configuration

The implementation follows the separated-data evaluation protocol proposed in the original work. Quantile regression models were trained using the public BloodMNIST dataset, while calibration and attack evaluation were performed using disjoint client datasets to avoid information leakage between the different attack stages.

Several confidence-based scoring functions were investigated, including hinge score, prediction confidence, loss, entropy and modified entropy. The hinge score was adopted as the primary attack score due to its superior performance during preliminary experiments.

The attack was executed against the converged global federated model under all investigated federated learning configurations, enabling the influence of Differential Privacy on membership leakage to be analysed.

The effectiveness of the Membership Inference Attack is quantitatively analysed in Section 3.8, where the corresponding privacy metrics are presented and discussed.

3.7.3 Frequency-Injection Backdoor Attack

The robustness of the proposed federated learning framework against data poisoning was evaluated using the Frequency-Injection Backdoor Attack (FIBA) proposed by Feng et al. [16]. Unlike conventional trigger-based backdoor attacks, FIBA embeds the malicious trigger in the frequency domain by modifying the Fourier spectrum of the training images. This approach produces visually subtle perturbations while remaining highly effective at manipulating the behaviour of the trained model.

The attack was selected because it represents a realistic threat in federated learning, where a malicious participant can locally poison its private dataset before contributing model updates to the collaborative training process.

Frequency-Domain Trigger Injection

The FIBA poisoning procedure operates on the frequency representation of an image. Given a benign training image and a predefined trigger image, the Fast Fourier Transform (FFT) is applied to obtain their corresponding frequency-domain representations. Information derived from the trigger is then injected into the frequency components of the benign sample, after which the modified representation is transformed back into the spatial domain using the inverse Fast Fourier Transform (iFFT).

This procedure allows the malicious trigger to be embedded without introducing an obvious spatial pattern in the poisoned image. Figure 3.5 illustrates the effect of the attack on a BloodMNIST sample. The top row compares the original image, the trigger image, and the resulting poisoned sample. The middle row shows their corresponding frequency spectra, while the bottom row highlights the differences introduced by the poisoning process in both the spatial and frequency domains. Although the poisoned image remains visually similar to the original sample, the injected modification becomes more evident when the corresponding frequency representations and difference maps are examined.

Integration into the Proposed Framework

The attack was integrated into the proposed Flower-based federated learning framework by simulating one malicious client among the three participating institutions. Before local training, the malicious client applied the FIBA poisoning procedure to a subset of its training images while simultaneously modifying their corresponding labels to a predefined target class.

The poisoning process was performed exclusively on the local dataset of the malicious client. Consequently, neither the central server nor the remaining honest clients had knowledge of the manipulated samples during collaborative training. After poisoning, the malicious client participated in the federated learning process following the same communication protocol as the benign clients, allowing the attack to remain indistinguishable from normal training behaviour.

This implementation reproduces the realistic threat scenario in which a compromised institution attempts to implant a hidden backdoor into the collaboratively trained global model without violating the federated learning protocol.

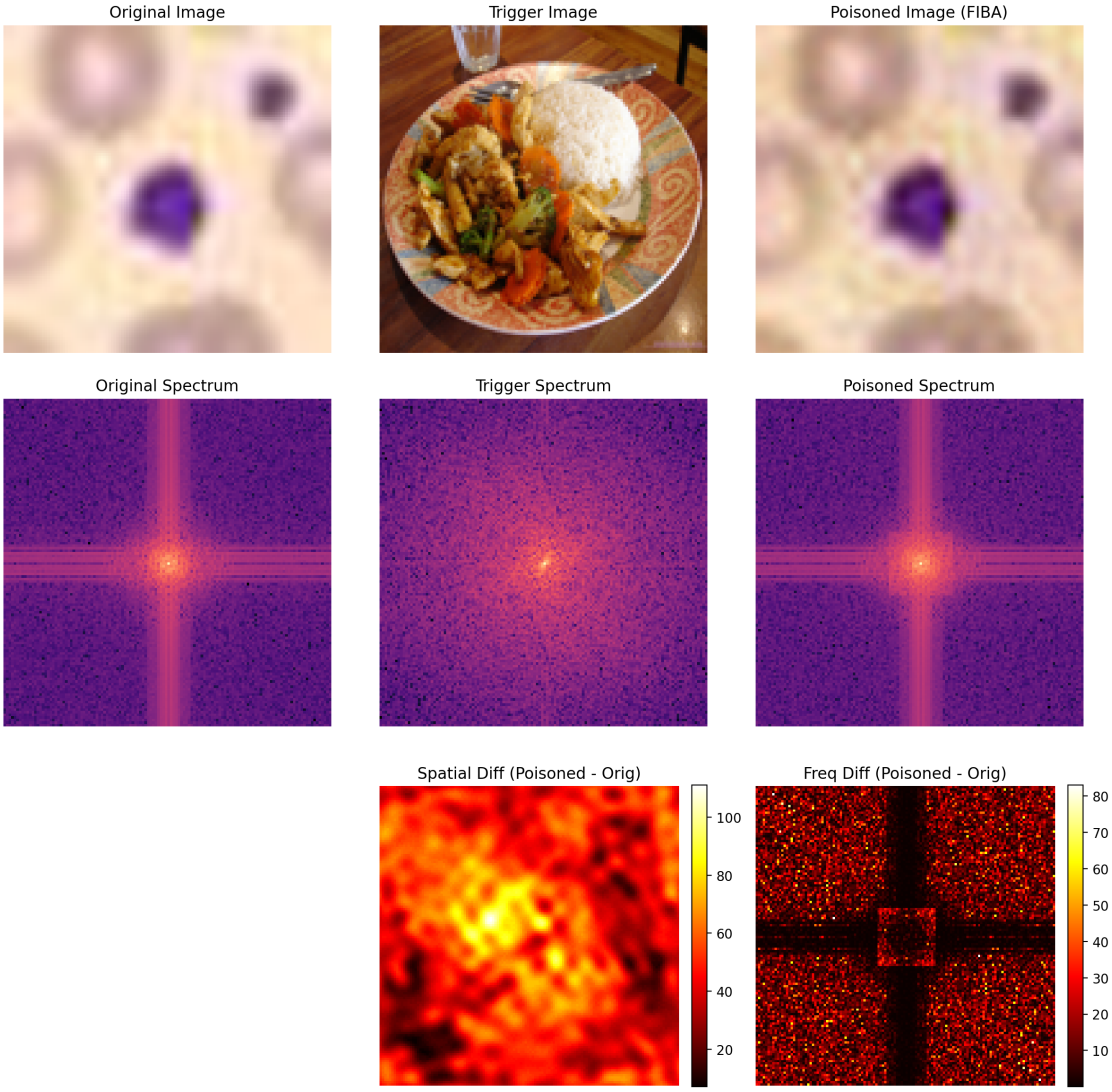


Figure 3.5: Visualization of the Frequency-Injection Backdoor Attack (FIBA). The top row shows the original BloodMNIST sample, the trigger image, and the resulting poisoned sample. The middle row presents their corresponding frequency spectra, while the bottom row highlights the differences between the original and poisoned samples in the spatial and frequency domains.

Experimental Configuration

In each backdoor experiment, one of the three participating clients was configured as malicious, while the remaining two clients behaved honestly. The attack targeted class 3 (*Immature Granulocytes*), meaning that the poisoned samples were relabelled as the target class during the malicious client's local training. The malicious client followed the same local training configuration as the benign clients, including the same number of local epochs.

The Frequency-Injection Backdoor Attack was evaluated under both FedAvg and CKADA aggregation, with and without Differential Privacy. This experimental design allows the influence of the aggregation strategy and Differential Privacy on the propagation of the backdoor through the federated training process to be investigated.

The effectiveness of the attack is evaluated in Section 3.8 using Benign Accuracy (BA) and Attack Success Rate (ASR), which respectively measure the performance of the global model on clean samples and its tendency to classify triggered samples as the predefined target class.

3.7.4 Gradient Inversion Attack

To evaluate the confidentiality of the information exchanged during federated learning, the proposed framework was assessed against a Gradient Inversion Attack (GIA). As described in Chapter 2, gradient inversion attacks attempt to reconstruct the original training samples from the gradients or model updates transmitted during collaborative optimization.

The implementation adopted in this work is based on the optimization-based attack proposed by Geiping et al. [19], which formulates image reconstruction as an iterative optimization problem that minimizes the distance between the observed client gradients and the gradients generated from reconstructed samples.

Integration into the Proposed Framework

The attack was executed after obtaining the local model updates generated by an individual federated learning client. These updates were treated as the information available to a curious adversary attempting to recover the client's private training data.

The implementation follows the optimization procedure proposed by Geiping et al., where dummy images are initialized randomly and iteratively optimized until the gradients produced by the reconstructed samples closely match the observed client gradients. The attack was performed using the same model architecture employed during federated training to ensure consistency between the optimization process and the victim model.

To evaluate the effectiveness of the proposed privacy-preserving framework, gradient inversion experiments were conducted using both protected and unprotected model updates, allowing the influence of Differential Privacy on reconstruction quality to be systematically analysed.

Experimental Configuration

The reconstruction process was performed using the original optimization strategy proposed by Geiping et al., including multiple optimization iterations and gradient matching across all trainable model parameters. The attack was applied to local model updates extracted from selected communication rounds after convergence of the federated learning process.

Separate experiments were conducted with and without Differential Privacy to quantify the extent to which gradient perturbation reduces the amount of recoverable information contained within the transmitted client updates. Homomorphic Encryption was not directly evaluated using the Gradient Inversion Attack, since encrypted model updates cannot be processed by the attacker without prior decryption.

The quality of the reconstructed images is evaluated in Section 3.8 using both pixel-wise and perceptual image quality metrics to quantify the effectiveness of the attack under different privacy-preserving configurations.

3.8 Security Evaluation

This section describes the methodology adopted to evaluate the security of the proposed privacy-preserving federated learning framework. Following the implementation of the adversarial scenarios presented in Section 3.7, the objective of the security evaluation is to quantify the framework's resistance against privacy leakage, model poisoning and data reconstruction attacks.

Each attack targets a different security property of the federated learning system. Consequently, dedicated evaluation protocols and performance metrics are employed to assess the effectiveness of the corresponding privacy-preserving mechanisms. The following subsections describe the evaluation methodology adopted for each attack, while the experimental results are presented and discussed in Chapter 4.

3.8.1 Evaluation Objectives

The security evaluation was designed to investigate three complementary aspects of the proposed framework.

First, the Membership Inference Attack evaluates whether the privacy-preserving mechanisms reduce the amount of information that can be inferred from the trained global model regarding the participation of individual training samples.

Second, the Frequency-Injection Backdoor Attack assesses the robustness of the aggregation process against malicious client updates intended to implant hidden behaviours into the global model.

Finally, the Gradient Inversion Attack evaluates whether sensitive training images can be reconstructed from the information exchanged during federated learning and analyses the effectiveness of Differential Privacy in mitigating this threat.

Together, these evaluations provide a comprehensive assessment of the confidentiality, integrity and privacy guarantees offered by the proposed framework.

3.8.2 Membership Inference Evaluation

The effectiveness of the Membership Inference Attack is evaluated by measuring the attacker's ability to distinguish between member and non-member samples using the prediction scores produced by the trained global model.

The attack is performed after the federated learning process has converged, using the separated-data evaluation protocol described in Section 3.7.2. Attack performance is quantified through threshold-independent and threshold-dependent metrics that measure the discriminative capability of the attack and its practical effectiveness.

Lower attack performance indicates that the trained model leaks less membership information and therefore provides stronger privacy protection.

3.8.3 Backdoor Evaluation

The robustness of the proposed framework against malicious client behaviour is evaluated using the Frequency-Injection Backdoor Attack introduced in Section 3.7.3.

The evaluation considers two complementary objectives. The first is to measure the classification performance of the global model on benign test samples, ensuring that the privacy-preserving mechanisms do not unnecessarily degrade predictive performance. The second is to quantify the success of the malicious trigger in forcing misclassification towards the target label.

An effective defence should preserve high classification accuracy on clean data while simultaneously reducing the effectiveness of the injected backdoor.

3.8.4 Gradient Inversion Evaluation

The Gradient Inversion Attack is evaluated by analysing the visual and numerical similarity between the reconstructed images and the original client training samples.

The attack is executed using the methodology described in Section 3.7.4 under different privacy-preserving configurations. Reconstruction quality is assessed through multiple complementary image quality metrics that quantify both pixel-level similarity and perceptual image fidelity.

Lower reconstruction quality indicates that less sensitive information can be recovered from the exchanged model updates, demonstrating improved resistance against gradient inversion attacks.

3.8.5 Security Metrics

Table 3.8 summarizes the quantitative metrics adopted to evaluate the effectiveness of the implemented attacks and the corresponding privacy-preserving mechanisms. Different metrics are used depending on the attack objective, allowing both the attack success and the effectiveness of the proposed defenses to be assessed from complementary perspectives.

The security evaluation methodology presented in this section establishes a consistent framework for analysing the effectiveness of the proposed privacy-preserving mechanisms against representative adversarial threats. The following section describes the complete

Table 3.8: Security metrics adopted for the evaluation of the implemented attacks.

Attack	**Metric**	**Evaluation**
Membership Inference	ROC AUC	Measures the overall ability of the attack to distinguish member from non-member samples. Higher values indicate stronger attacks.
	Average Precision	Summarizes the precision–recall curve, providing a robust evaluation under class imbalance. Higher values indicate better attack performance.
	TPR@FPR	Measures the true positive rate at fixed false positive rate thresholds, evaluating attack effectiveness in security-critical operating regions.
Frequency-Injection Backdoor	Benign Accuracy	Classification accuracy on clean, non-poisoned test samples, measuring the preservation of normal model performance.
	Attack Success Rate (ASR)	Percentage of triggered inputs classified as the attacker-selected target class. Higher values indicate more successful backdoor attacks.
Gradient Inversion	MSE	Average pixel-wise reconstruction error between the reconstructed and original images. Lower values indicate better reconstruction.
	PSNR	Peak Signal-to-Noise Ratio measuring reconstruction fidelity. Higher values indicate better image quality.
	MS-SSIM	Measures structural similarity between reconstructed and original images across multiple scales. Higher values indicate better structural preservation.
	UQI	Evaluates overall image similarity considering luminance, contrast, and structural information. Higher values indicate greater similarity.
	VIF	Measures the amount of visual information preserved in the reconstructed image relative to the original. Higher values indicate better visual fidelity.
	FMSE	Mean squared error computed in feature space rather than pixel space, measuring semantic similarity. Lower values indicate better reconstruction.

experimental campaign conducted to evaluate the proposed framework under different federated learning configurations before presenting the experimental results in Chapter 4.

3.9 Experimental Design

This section describes the experimental campaign conducted to evaluate the proposed privacy-preserving federated learning framework. The experiments were designed to analyse the impact of different aggregation strategies, privacy-preserving mechanisms, and data distributions on model performance, computational efficiency, and security. In addition, a centralized learning approach was implemented as a baseline to quantify the benefits and trade-offs introduced by federated learning.

3.9.1 Experimental Objectives

The experimental campaign was designed to address the research questions presented in Section 3.1. The experiments first compare the predictive performance of centralized and federated learning, providing a reference for evaluating the effect of collaborative training. They then investigate the influence of different aggregation strategies under varying data distributions in order to assess the impact of statistical heterogeneity on federated model performance.

A further objective is to analyse the effect of Differential Privacy on model utility and on resistance to privacy attacks. Homomorphic Encryption is evaluated separately to quantify the additional system overhead introduced by encrypted aggregation, with particular attention to communication cost and computational considerations. Finally, the experimental campaign assesses the security of the federated learning framework against representative threats, including Membership Inference, Gradient Inversion, and Backdoor attacks.

All experiments were conducted using the same model architecture, training configuration, and evaluation protocol to ensure a fair comparison between the investigated scenarios.

3.9.2 Baseline Comparison

The first experiment compares the proposed federated learning framework with a conventional centralized learning approach. Both models were trained using the same dataset, neural network architecture and optimization parameters.

This comparison establishes a reference point for evaluating the impact of distributed training on predictive performance, computational cost and communication overhead.

3.9.3 Aggregation Strategy Evaluation

The second group of experiments investigates the influence of the aggregation strategy on the performance of federated learning. Two aggregation algorithms were evaluated: Federated Averaging (FedAvg) and CKADA. The complete set of evaluated configurations is summarized in Table 3.9.

Each aggregation strategy was tested under three different client data distributions. This experimental design enables the analysis of aggregation robustness under increasing levels of statistical heterogeneity.

Table 3.9: Aggregation strategy evaluation scenarios.

Experiment	Aggregation	Data Distribution
E1	FedAvg	IID
E2	FedAvg	Dirichlet ($\alpha = 0.5$)
E3	FedAvg	Dirichlet ($\alpha = 0.1$)
E4	CKADA	IID
E5	CKADA	Dirichlet ($\alpha = 0.5$)
E6	CKADA	Dirichlet ($\alpha = 0.1$)

3.9.4 Differential Privacy Evaluation

The third group of experiments evaluates the effect of Differential Privacy on both model utility and privacy protection. As summarized in Table 3.10, different privacy budgets were considered to analyse the trade-off between predictive performance and resistance against privacy attacks.

Each Differential Privacy configuration was evaluated using both aggregation strategies under the same experimental conditions adopted for the baseline federated learning experiments.

Table 3.10: Differential Privacy evaluation scenarios.

Experiment	Aggregation	Privacy Budget (ε)
E7	FedAvg	10
E8	FedAvg	3
E9	FedAvg	1
E10	CKADA	10
E11	CKADA	3
E12	CKADA	1

3.9.5 Homomorphic Encryption Evaluation

The fourth group of experiments investigates the computational overhead introduced by Homomorphic Encryption. The evaluated configurations are summarized in Table 3.11. Since encrypting all model parameters substantially increases communication and computation costs, different encryption ratios were evaluated to analyse the trade-off between privacy protection and system efficiency.

Only the selected proportion of model layers specified by the encryption ratio was protected using the CKKS encryption scheme.

Table 3.11: Homomorphic Encryption evaluation scenarios.

Experiment	**Aggregation**	**Encryption Ratio**
E13	FedAvg	25%
E14	FedAvg	50%
E15	FedAvg	100%
E16	Hybrid	25%
E17	Hybrid	50%
E18	Hybrid	100%

Chapter 4

Results and Discussion

4.1 Experimental Overview

This chapter presents the experimental evaluation of the proposed privacy-preserving federated learning framework. The objective of the experimental campaign is to assess the effectiveness of the proposed solution from three complementary perspectives: predictive performance, computational efficiency, and resistance against privacy and security attacks.

The evaluation is organized into a sequence of experiments that progressively investigate different aspects of the proposed framework. As summarized in Table 4.1, the experimental campaign comprises three main groups of experiments: baseline federated learning configurations, Differential Privacy enabled configurations, and Homomorphic Encryption enabled configurations. First, the performance of Federated Learning is compared with that of conventional centralized learning to quantify the impact of distributed training on classification accuracy, convergence behaviour, and computational cost. Subsequently, the influence of data heterogeneity is analysed by considering both IID and non-IID client distributions together with different aggregation strategies.

The second part of the experimental campaign evaluates the proposed privacy-preserving mechanisms. Differential Privacy and Homomorphic Encryption are analysed independently to quantify their impact on model utility, computational overhead, and communication cost. Finally, the complete framework is evaluated under representative privacy and security attacks, including Membership Inference, Gradient Inversion, and Frequency-Injection Backdoor attacks, to assess the effectiveness of the proposed protection mechanisms.

Unless otherwise stated, all experiments were conducted using the BloodMNIST dataset described in Chapter 3.3, the ResNet-18 architecture presented in Section 3.4, and the Flower-based federated learning framework introduced in Section 3.5. The training configuration and evaluation metrics remained unchanged throughout the experiments to ensure fair and reproducible comparisons between all evaluated scenarios.

Table 4.1: Summary of the experimental configurations evaluated in this work.

Exp.	**Objective**	**Aggregation**	**DP**	**HE**	**Distribution**
E1	Baseline FL	FedAvg	–	–	IID
E2	Baseline FL	FedAvg	–	–	Dirichlet ($\alpha = 0.5$)
E3	Baseline FL	FedAvg	–	–	Dirichlet ($\alpha = 0.1$)
E4	Baseline FL	CKADA	–	–	IID
E5	Baseline FL	CKADA	–	–	Dirichlet ($\alpha = 0.5$)
E6	Baseline FL	CKADA	–	–	Dirichlet ($\alpha = 0.1$)
E7	DP Evaluation	FedAvg	$\varepsilon = 10$	–	IID
E8	DP Evaluation	FedAvg	$\varepsilon = 3$	–	IID
E9	DP Evaluation	FedAvg	$\varepsilon = 1$	–	IID
E10	DP Evaluation	CKADA	$\varepsilon = 10$	–	IID
E11	DP Evaluation	CKADA	$\varepsilon = 3$	–	IID
E12	DP Evaluation	CKADA	$\varepsilon = 1$	–	IID
E13	HE Evaluation	FedAvg	–	25%	IID
E14	HE Evaluation	FedAvg	–	50%	IID
E15	HE Evaluation	FedAvg	–	100%	IID
E16	HE Evaluation	Hybrid	–	25%	IID
E17	HE Evaluation	Hybrid	–	50%	IID
E18	HE Evaluation	Hybrid	–	100%	IID

4.2 Centralized vs Federated Learning

The first experimental comparison investigates the differences between conventional centralized learning and the baseline federated learning configuration. The centralized model was trained using the complete BloodMNIST training dataset at a single location, whereas the federated model was trained across three clients using the FedAvg aggregation algorithm and an IID data distribution.

The federated configuration corresponds to experiment E1. Differential Privacy and Homomorphic Encryption were disabled in this experiment so that the comparison would capture the intrinsic effects of distributed training without additional perturbation or cryptographic overhead. Both configurations employed the same ResNet-18 architecture, data preprocessing pipeline, optimization algorithm, and central evaluation dataset.

The comparison considers predictive performance, execution time, computational cost, and communication overhead. For each learning approach, the checkpoint achieving the highest central test accuracy is used for the predictive comparison. The centralized model reached its best accuracy at epoch 32, while the federated model reached its highest global accuracy at communication round 16.

4.2.1 Performance Comparison

Table 4.2 compares the predictive performance of the centralized model and the E1 federated model at their respective best checkpoints.

Table 4.2: Predictive performance of centralized learning and federated learning under the IID configuration.

Metric	**Centralized**	**Federated E1**	**Difference (Δ)**
Accuracy	0.9418	0.9187	-0.0231
F1-score	0.9345	0.9075	-0.027
AUC	0.9955	0.9926	-0.0029
Loss	0.3449	0.2413	-0.1036

The centralized model achieved the highest classification accuracy, exceeding the federated model by approximately 2.3 percentage points. A similar difference was observed for the F1-score, for which the centralized and federated configurations achieved values of 0.9345 and 0.9075, respectively. The F1-score difference of 0.027 indicates that the centralized model provided a slightly better balance between precision and recall across the eight BloodMNIST classes.

The difference in AUC was considerably smaller. The centralized model achieved an AUC of 0.9955, compared with 0.9926 for the federated model. Both values are close to one, indicating that both models retained a strong ability to distinguish between the different blood-cell classes. Therefore, although centralized learning achieved the best overall performance, the federated model preserved most of the predictive capability of the centralized baseline.

Interestingly, the federated model reported a lower cross-entropy loss at its best-accuracy checkpoint. Its loss was 0.2413, compared with 0.3449 for centralized learning. This result does not contradict the lower federated accuracy. Accuracy considers only whether the predicted class is correct, whereas cross-entropy loss also depends on the probability assigned to each class. Consequently, the federated model may produce better-calibrated or less heavily penalized probability distributions while still misclassifying slightly more samples. For this reason, loss should be interpreted together with accuracy, F1-score, and AUC rather than as an isolated indicator of model quality.

Overall, the reduction in predictive performance introduced by federated learning was limited. These results indicate that the E1 configuration can train an effective global model without centralizing the clients' healthcare data.

4.2.2 Computational Analysis

In addition to predictive performance, the computational cost of centralized and federated learning was evaluated in terms of execution time and floating-point operations (FLOPs). To ensure a fair comparison, all computational metrics are reported up to the point of convergence rather than for the complete training execution.

In this work, convergence is defined as the first optimization iteration (training epoch for centralized learning or communication round for federated learning) at which the classification accuracy reaches at least 99% of the highest accuracy observed during the entire training process. This criterion provides a consistent estimate of the computational resources required to obtain a near-optimal model while avoiding unnecessary computation after performance has effectively stabilized.

Although training continues until the early stopping criterion is satisfied, using a patience of 50 consecutive iterations without improvement in the validation accuracy, the computational analysis presented in this section considers only the resources consumed up to the convergence point. Consequently, the reported execution time and FLOPs reflect the cost required to achieve a stable predictive performance rather than the total cost of the complete training procedure. The computational results reported in Table 4.3 show

Table 4.3: Computational comparison between centralized learning and the E1 federated configuration until convergence.

Metric	**Centralized**	**Federated E1**	**Relative Factor**
Convergence epoch / round	32	16	N/A
Execution time (s)	1,558	6,434.77	4.13×
Training FLOPs	8.39×10^{14}	4.19×10^{15}	5.00×
Evaluation FLOPs	1.00×10^{14}	5.31×10^{13}	0.53×
Total FLOPs	9.39×10^{14}	4.25×10^{15}	4.52×

that federated learning requires greater computational resources than centralized learning, even when both approaches are evaluated only until convergence. The centralized model reached convergence after 32 training epochs, whereas the federated configuration converged after 16 communication rounds.

The difference is particularly evident in the execution time. Although the federated model converged in fewer optimization iterations, it required approximately 6,435 seconds to reach convergence, compared with 1,558 seconds for centralized learning, corresponding to a relative execution time of approximately 4.13×. This behaviour is expected because each communication round involves multiple stages, including local training at each participating client, transmission of model updates to the server, aggregation of the received parameters, and redistribution of the updated global model.

A similar trend is observed in the computational workload. Centralized learning required approximately 8.39×10^{14} training FLOPs to reach convergence, whereas the federated configuration required approximately 4.19×10^{15} training FLOPs, corresponding to a relative factor of approximately 5.00×. This increase is mainly attributable to the distributed nature of federated learning, where all participating clients independently perform local optimization during every communication round before their updates are aggregated by the server.

Interestingly, the cumulative evaluation cost was lower for the federated configuration. The centralized model required approximately 1.00×10^{14} evaluation FLOPs, whereas the federated model required only 5.31×10^{13} evaluation FLOPs, corresponding to a relative factor of 0.53×. This reduction is explained by the earlier convergence of the federated model, which required fewer global evaluations before reaching the convergence criterion.

Overall, the total computational cost of federated learning remained approximately 4.52× higher than that of centralized learning. Although this represents a significant computational overhead, it reflects the additional operations required to perform collaborative optimization while maintaining data locality, since no raw medical images are exchanged between participating healthcare institutions.

4.2.3 Communication Analysis

Unlike centralized learning, federated learning requires repeated communication between the participating clients and the central server. Table 4.4 summarizes the communication statistics measured for the E1 configuration until convergence.

Table 4.4: Communication metrics of the E1 federated configuration until convergence.

Metric	**Value**
Communication rounds	16
Client upload per round	44.69 MB
Server download per round	134.10 MB
Server upload per round	268.17 MB

During each communication round, every client transmitted approximately 44.69 MB of model parameters to the server, resulting in a cumulative server download of approximately 134.10 MB from the three participating clients. After aggregating the received updates, the server transmitted the global model twice to all clients: first to initiate the

next round of local training and subsequently to enable local model evaluation. Consequently, the cumulative server upload reached approximately 268.17 MB per communication round. Although no raw medical data were exchanged, the repeated transmission of model parameters constitutes an inherent communication overhead of federated learning and contributes to the overall execution time.

4.2.4 Discussion

The comparison between centralized and federated learning highlights the fundamental trade-off between predictive performance and data locality. As expected, centralized learning achieved the highest classification performance by optimizing a single model using the complete training dataset. Nevertheless, the reduction in predictive performance introduced by federated learning remained limited. The E1 configuration (Table 4.1) achieved an accuracy and AUC comparable to the centralized baseline, demonstrating that collaborative training can preserve most of the predictive capability of centralized learning while maintaining the decentralization of sensitive healthcare data.

From a computational perspective, federated learning required greater resources than centralized training. Although the federated configuration converged after only 16 communication rounds, compared with 32 training epochs for the centralized model, each communication round includes local optimization on all participating clients together with the transmission and aggregation of model updates. Consequently, the federated configuration required approximately 4.13× longer execution time and 5.00× more training FLOPs, resulting in an overall computational cost approximately 4.52× higher than centralized learning.

Communication also represents a significant source of overhead in federated learning. During each communication round, every client uploaded approximately 44.7 MB of model parameters to the server, while the server transmitted approximately 268.2 MB of model data to coordinate training and evaluation across the three participating clients. Although no raw medical images were exchanged, the repeated communication of model parameters contributes directly to the overall execution time and may become a limiting factor as the number of participating institutions, communication rounds, or model size increases.

Despite these computational and communication overheads, federated learning offers an important advantage for healthcare applications by eliminating the need to centralize sensitive patient data. Each participating institution retains its local dataset throughout the entire training process, while only model updates are exchanged with the coordinating server. The additional resource requirements should therefore be interpreted as the cost of enabling collaborative learning while preserving data locality.

It is important to emphasize that preserving data locality alone does not guarantee complete privacy. Although raw medical images are never transmitted, the exchanged model updates may still reveal information about the underlying training data through attacks such as Membership Inference and Gradient Inversion. Consequently, the E1 configuration establishes the baseline against which the proposed privacy-preserving mechanisms are evaluated in the following sections. The subsequent experiments investigate how Differential Privacy, Homomorphic Encryption, and their combination influence model

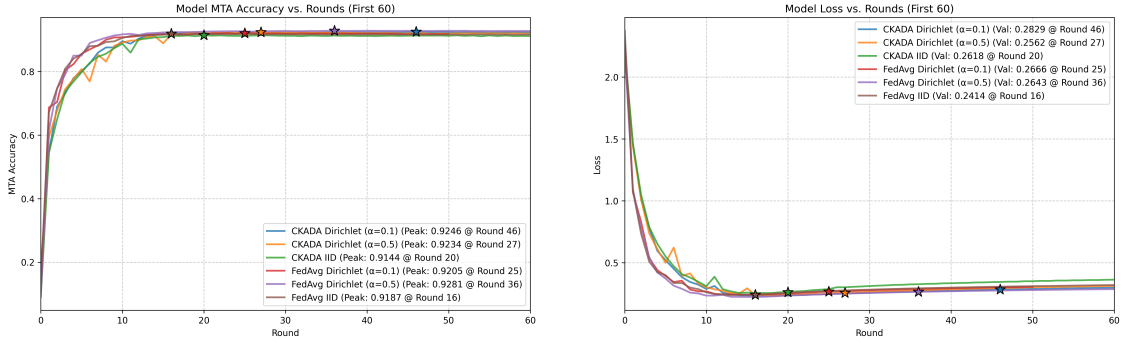
utility, computational efficiency, and resistance to representative privacy and security attacks.

Overall, the results demonstrate that federated learning provides a practical alternative to centralized learning for collaborative healthcare applications. While additional computational resources are required, the framework achieves comparable predictive performance without requiring healthcare institutions to share their raw patient data, thereby establishing a suitable foundation for the privacy-preserving mechanisms proposed in this thesis.

4.3 Impact of Data Distribution and Aggregation

This section investigates the influence of client data heterogeneity and aggregation strategy on the performance of the proposed federated learning framework. Six baseline configurations were evaluated by combining two aggregation algorithms (FedAvg and CKADA) with three data partitioning strategies, namely IID, moderate non-IID ($\alpha = 0.5$), and severe non-IID ($\alpha = 0.1$).

Figures 4.1a–4.1b illustrate the evolution of the principal evaluation metrics throughout the first 60 communication rounds. The stars indicate the communication round at which each configuration achieved its highest classification accuracy.



(a) Classification accuracy

(b) Training loss

Figure 4.1: Comparison of the effects of aggregation strategy (FedAvg and CKADA) and client data distribution (IID, Dirichlet $\alpha = 0.5$, and Dirichlet $\alpha = 0.1$) on model performance.

4.3.1 IID Distribution

Under the IID partitioning strategy, the training data were distributed approximately uniformly across the three participating clients, resulting in similar local class distributions. This configuration represents the ideal federated learning scenario, in which statistical heterogeneity between institutions is minimized.

As shown in Figures 4.1a and 4.1b, both aggregation strategies rapidly converged to comparable predictive performance. FedAvg achieved a peak accuracy of 0.9187 after

only 16 communication rounds, whereas CKADA reached a similar peak accuracy of 0.9144 after 20 rounds. The corresponding AUC values remained close to 0.99 throughout training (Appendix, Figure A.1a), indicating excellent discriminative capability for both methods.

Although the final predictive performance differed by less than one percentage point, FedAvg converged more rapidly than CKADA. The loss curves exhibit a steep decrease during the first ten communication rounds, followed by a stable convergence phase with only minor fluctuations. These results indicate that, under homogeneous data distributions, both aggregation strategies achieve stable optimization, with FedAvg providing slightly faster convergence.

4.3.2 Moderate Non-IID ($\alpha = 0.5$)

A moderate degree of statistical heterogeneity was introduced by partitioning the training data according to a Dirichlet distribution with concentration parameter $\alpha = 0.5$. Compared with the IID configuration, clients therefore observed different class proportions while still maintaining sufficient class diversity for local optimization.

This configuration produced the highest predictive performance among all evaluated scenarios. FedAvg achieved the highest overall accuracy (0.9281), converging after 36 communication rounds, while CKADA reached a very similar accuracy of 0.9234 after only 27 rounds. The corresponding loss curves remained stable after convergence, indicating that both aggregation strategies effectively handled the moderate level of statistical heterogeneity.

These results indicate that moderate non-IID data distributions can be effectively handled by both aggregation algorithms. A contributing factor is that the proposed framework does not employ client selection; instead, all participating clients contribute to every communication round. Consequently, the global model is updated using information from the complete distributed dataset at each aggregation step. Although the local class distributions differ across clients, every class remains represented during global optimization, allowing the model to maintain high predictive performance despite the presence of statistical heterogeneity.

4.3.3 Severe Non-IID ($\alpha = 0.1$)

The most challenging experimental scenario was obtained by partitioning the training data according to a Dirichlet distribution with concentration parameter $\alpha = 0.1$, resulting in highly heterogeneous client datasets in which individual classes became concentrated within a limited number of participating institutions.

Despite the increased statistical heterogeneity, both aggregation strategies remained robust. FedAvg achieved a peak accuracy of 0.9205 after 25 communication rounds, while CKADA obtained a slightly higher accuracy of 0.9246 after 46 rounds. Although CKADA ultimately achieved the highest accuracy in this setting, it required almost twice as many communication rounds to reach convergence. The corresponding loss curves remained stable after convergence, suggesting that neither aggregation strategy experienced significant optimization difficulties.

The limited performance degradation can be attributed, in part, to the federated training protocol adopted in the proposed framework (Section 3.5). Unlike many practical federated learning deployments that rely on client sampling, all three clients participate in every communication round in the present implementation. Consequently, the server aggregates model updates computed from the complete distributed dataset rather than from only a subset of clients. Although the local class distributions are highly skewed, every class contributes to each global model update, preventing the loss of class-specific information that may otherwise occur when only a fraction of clients is selected for training. This design choice mitigates the negative impact of severe statistical heterogeneity and explains the consistently high predictive performance observed for both aggregation strategies.

4.3.4 FedAvg vs CKADA

Across all three partitioning strategies, FedAvg and CKADA achieved comparable predictive performance. The maximum accuracy differed by less than approximately one percentage point in every configuration, while the corresponding AUC values consistently remained close to 0.99.

The principal difference between the two aggregation strategies lies in their convergence behaviour rather than their final predictive performance. FedAvg consistently reached its optimum using fewer communication rounds, converging after 16, 25, and 36 rounds for the IID, $\alpha = 0.1$, and $\alpha = 0.5$ configurations, respectively. CKADA required 20, 27, and 46 communication rounds under the same settings, indicating a slower optimization process.

Despite requiring additional communication rounds, CKADA achieved predictive performance comparable to FedAvg and even produced the highest accuracy under the severe non-IID configuration. These results suggest that, for BloodMNIST, the choice between FedAvg and CKADA has only a limited impact on model utility. Instead, the decision should primarily be guided by other considerations, such as robustness against adversarial attacks, which is investigated in the security evaluation presented later in this chapter.

4.4 Impact of Differential Privacy

This section evaluates the impact of Differential Privacy (DP) on the proposed federated learning framework. The analysis focuses on three complementary aspects: the influence of DP on predictive performance, its effectiveness in mitigating privacy attacks, and the computational overhead introduced during the training process. The objective is to quantify the trade-off between privacy protection and model utility under different privacy budgets.

4.4.1 Utility Analysis

To assess the effect of Differential Privacy on predictive performance, the proposed framework was evaluated under multiple privacy budgets and compared with the corresponding

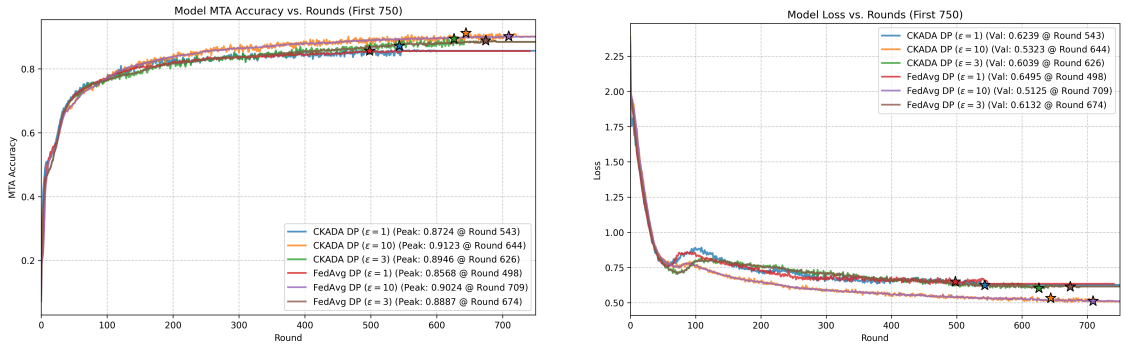
non-private federated learning configuration. Model utility was assessed using classification accuracy, F1-score, AUC, and validation loss.

Table 4.5 summarizes the predictive performance obtained for the evaluated privacy budgets, while Figures 4.2a–4.2b illustrate the corresponding convergence behaviour throughout the federated training process.

Table 4.5: Predictive performance obtained under different Differential Privacy configurations.

Configuration	**Aggregation**	**Accuracy**	**F1-score**	**AUC**	**Loss**
No DP	FedAvg	0.9187	0.9075	0.9925	0.2413
$\varepsilon = 1$	FedAvg	0.8567	0.8064	0.9764	0.6494
$\varepsilon = 3$	FedAvg	0.8839	0.8592	0.9845	0.6074
$\varepsilon = 10$	FedAvg	0.9023	0.8819	0.9886	0.5125
No DP	CKADA	0.9143	0.9021	0.9917	0.2618
$\varepsilon = 1$	CKADA	0.8724	0.7952	0.9855	0.6238
$\varepsilon = 3$	CKADA	0.8946	0.8516	0.9891	0.6038
$\varepsilon = 10$	CKADA	0.9123	0.8898	0.9874	0.5322

As expected, increasing the strength of Differential Privacy introduces progressively larger perturbations into the optimization process. Consequently, stronger privacy guarantees are generally accompanied by a gradual reduction in predictive performance and slower convergence. Nevertheless, the results demonstrate whether acceptable classification performance can still be maintained while enforcing formal privacy guarantees.



(a) Classification accuracy

(b) Training loss

Figure 4.2: Comparison of the effects of aggregation strategy (FedAvg and CKADA) and DP privacy budget ($\varepsilon = 1$, $\varepsilon = 3$, $\varepsilon = 10$) on model performance.

4.4.2 Computational Overhead

Although Differential Privacy improves privacy protection, it introduces additional runtime overhead during local model optimization. Unlike Homomorphic Encryption, Differential Privacy does not change the neural network architecture or the number of

arithmetic operations required for model training. Consequently, the estimated training FLOPs remain unchanged. The observed overhead instead originates from the additional processing required by the Differential Privacy algorithm, including per-sample gradient computation, gradient clipping, Gaussian noise injection, and privacy accounting.

Table 4.6 compares the runtime characteristics of the federated learning framework with and without Differential Privacy. Since the experiments may require different numbers of communication rounds, the computational overhead is evaluated using the average execution time per training round rather than the cumulative execution time.

Table 4.6: Average execution time with and without Differential Privacy.

Metric	**Without DP**	**With DP**	**Relative Factor**
Average execution time per round (s)	402.17	889.66	2.21×

4.4.3 Discussion

The results demonstrate that Differential Privacy introduces a substantial computational overhead during federated training. The average execution time per communication round increases from 402.17s without DP to 889.66s with DP, corresponding to a relative factor of approximately 2.21 times. This increase is primarily associated with the additional operations required by DP-SGD, including per-sample gradient processing, gradient clipping, and noise injection performed through Opacus. These operations increase the practical training cost even though the underlying model architecture remains unchanged.

This computational overhead represents an important trade-off for the adoption of Differential Privacy in federated healthcare applications. While DP increases training time, it provides an additional layer of protection against privacy attacks by limiting the information that can be inferred from individual client updates. Therefore, the additional computational cost should be considered together with the privacy improvements obtained during the security evaluation. Overall, the results indicate that Differential Privacy can strengthen the privacy guarantees of the federated learning framework, but this protection is achieved at the cost of increased execution time and, depending on the selected privacy budget, a potential reduction in predictive performance. The choice of the privacy parameter ε should consequently balance privacy protection, model utility, and computational efficiency according to the requirements of the target healthcare scenario.

4.5 Impact of Homomorphic Encryption

This section evaluates the impact of Homomorphic Encryption (HE) on the proposed federated learning framework. Unlike Differential Privacy, which modifies the optimization process by perturbing model updates, Homomorphic Encryption preserves the original model parameters while protecting them during transmission and aggregation. Consequently, the analysis focuses on three aspects: predictive performance, communication overhead, and computational cost introduced by encrypted aggregation.

4.5.1 Model Performance

Since Homomorphic Encryption operates exclusively on the communication process and does not alter the optimization algorithm, it is expected to preserve the predictive performance of the trained model. To verify this hypothesis, multiple encryption configurations were evaluated by varying the proportion of encrypted model layers through the `he_ratio` parameter.

Table 4.7 summarizes the predictive performance obtained under the evaluated encryption configurations.

Table 4.7: Predictive performance under different Homomorphic Encryption configurations.

HE Ratio	Aggregation	Accuracy	F1-score	AUC	Loss
0%	FedAvg	0.9187	0.9075	0.9925	0.2413
25%	FedAvg	0.9199	0.9084	0.9926	0.2909
50%	FedAvg	0.9213	0.9109	0.9930	0.2412
100%	FedAvg	0.9208	0.9090	0.9926	0.2830
0%	CKADA	0.9143	0.9021	0.9917	0.2618
25%	Hybrid	0.9266	0.9156	0.9931	0.2443
50%	Hybrid	0.9172	0.9057	0.9922	0.2797
100%	Hybrid	0.9099	0.8959	0.9913	0.2875

The results presented in Table 4.7 allow the influence of encrypted aggregation on model utility to be quantified. Since Homomorphic Encryption protects the exchanged model parameters without modifying their numerical values, only negligible differences in predictive performance are expected across the evaluated encryption levels. Any observed variations are therefore primarily attributable to the approximate arithmetic of the CKKS scheme or numerical precision effects rather than to changes in the learning process itself.

4.5.2 Communication Overhead

Although Homomorphic Encryption preserves model utility, it increases the size of the transmitted model updates because encrypted parameters occupy substantially more memory than their plaintext counterparts. Furthermore, the communication overhead depends on the proportion of encrypted layers defined by the `he_ratio` parameter.

Table 4.8 compares the communication cost associated with the evaluated encryption configurations.

Total server traffic is computed as the sum of server upload and server download during one communication round. As the encryption ratio increases, a larger proportion of the model parameters must be serialized, encrypted, transmitted, and aggregated in ciphertext form. Consequently, higher encryption ratios are expected to incur progressively greater communication overhead. This experiment therefore evaluates the trade-off between communication efficiency and the level of cryptographic protection achieved by the proposed selective encryption strategy.

Table 4.8: Communication overhead introduced by different Homomorphic Encryption ratios.

HE Ratio	Client Upload (MB) per round	Server Traffic (MB) per round	Relative Factor
0%	44.70	402	–
25%	57.27	465	1.16×
50%	70.82	532	1.32×
100%	924.96	4728	11.75×

4.5.3 Computational Considerations

Homomorphic Encryption introduces additional computational operations during the communication and aggregation phases of federated learning. In the proposed framework, local model training is performed in plaintext, while encryption is applied only to the selected model parameters after local optimization. The server subsequently performs aggregation on the encrypted parameters before the resulting values are decrypted.

Since the forward pass, backward pass, and local optimization procedure remain unchanged, the number of training FLOPs is effectively identical across the evaluated HE configurations. Consequently, conventional FLOP-based metrics do not capture the computational cost of cryptographic operations such as ciphertext encoding, encryption, homomorphic aggregation, and decryption.

Similarly, the total execution time required to reach convergence was not considered a direct measure of the overhead introduced by HE. Different configurations reached convergence after different numbers of communication rounds, meaning that their cumulative execution times reflect both cryptographic processing and differences in convergence behaviour. A direct comparison of total runtime would therefore risk attributing convergence-related variations to the encryption mechanism itself.

For this reason, the overhead introduced by HE is primarily evaluated through the communication analysis presented in Table4.8. The increase in transmitted data provides a direct and measurable consequence of encrypting a larger proportion of model parameters. In particular, the substantial increase observed under full encryption highlights the cost of applying CKKS to the complete model, whereas selective encryption limits this overhead by protecting only a subset of the exchanged parameters.

4.5.4 Discussion

The experimental results indicate that Homomorphic Encryption can be integrated into the federated learning framework without substantially affecting model utility. Across the evaluated FedAvg configurations, accuracy, F1-score, and AUC remain comparable to the unencrypted baseline, confirming that protecting model parameters during communication does not significantly interfere with the underlying optimization process. Similar behaviour is observed for the hybrid aggregation configurations, although small variations remain due to the interaction between the aggregation strategy and the proportion of encrypted layers.

The main cost associated with HE is instead observed in communication. Increasing the encryption ratio progressively increases the amount of data exchanged between clients and the server. This effect becomes particularly pronounced under full encryption, where total server traffic increases from 402 MB to 4728 MB per communication round, corresponding to approximately 11.75 times the communication cost of the unencrypted baseline. In contrast, selective encryption at 25% and 50% introduces substantially lower communication overhead, with relative factors of only 1.16 times and 1.32 times, respectively.

From a computational perspective, HE does not modify the local neural network training procedure and therefore does not increase the training FLOPs of the model. Nevertheless, cryptographic operations introduce additional processing that is not represented by conventional FLOP measurements. Moreover, cumulative execution time cannot be used to isolate this overhead reliably because the evaluated configurations converge after different numbers of communication rounds.

Overall, these findings support the use of selective Homomorphic Encryption as a practical compromise between cryptographic protection and system efficiency. While full encryption provides protection for all transmitted model parameters, it introduces a substantial communication burden. Selective encryption can instead protect the most relevant portions of the model while maintaining predictive performance and significantly reducing the additional communication cost, making it more suitable for resource-constrained federated healthcare environments.

4.6 Security Evaluation

4.6.1 Quantile Membership Inference Attack

To evaluate the privacy of the trained federated models, a Quantile Membership Inference Attack (Quantile MIA) was performed using the hinge confidence score. The attack attempts to determine whether a given sample was part of the training set by exploiting differences in the model's behaviour on member and non-member samples.

Four configurations were evaluated: FedAvg, FedAvg combined with Differential Privacy with three different privacy budgets $\varepsilon = 1$, $\varepsilon = 3$, $\varepsilon = 10$. The attack was performed on the converged global model obtained for each configuration using the experimental setup described in Section 3.8. Performance was assessed using the Area Under the ROC Curve (AUC), together with true positive rates measured at several fixed false positive rate thresholds.

The convergence round is also reported in Table 4.9 to identify the specific global model checkpoint used for each attack evaluation. Since Differential Privacy affects the optimization process and may require a different number of communication rounds to reach convergence, reporting this value ensures that the membership inference results are associated with the final converged model of each configuration rather than with models obtained after an identical number of training rounds. The reported round should therefore be interpreted as an experimental reference and not as a metric of attack effectiveness. The complete ROC curves on a logarithmic scale are reported in Appendix A, Figure A.3.

Discussion

The results reported in Table 4.9 show that the Quantile Membership Inference Attack has limited effectiveness against the evaluated federated learning configurations. In the baseline configuration, the attack achieves a ROC AUC of 0.5531, indicating only a weak ability to distinguish between member and non-member samples. Although this value

Table 4.9: Quantile Membership Inference Attack performance across the evaluated federated learning configurations.

Configuration	**Round**	**ROC AUC**	**AP**	**TPR@10% FPR**	**TPR@1% FPR**
Baseline (No DP)	16	0.5531	0.8159	0.0991	0.0147
DP with $\varepsilon = 1$	498	0.5006	0.7990	0.0806	0.0154
DP with $\varepsilon = 3$	630	0.5132	0.8076	0.1116	0.0088
DP with $\varepsilon = 10$	709	0.5193	0.8048	0.0903	0.0097

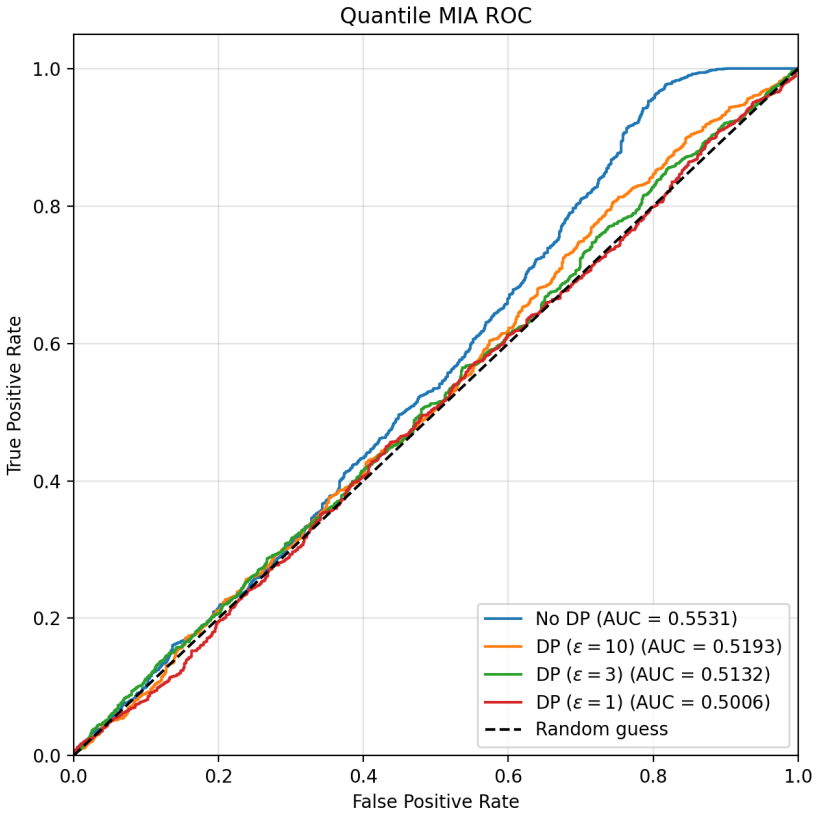


Figure 4.3: ROC curves of the Quantile Membership Inference Attack for the evaluated federated learning configurations.

is slightly above the 0.5 level corresponding to random guessing, it suggests that the baseline model exposes only a limited amount of exploitable membership information.

The introduction of Differential Privacy further reduces the overall effectiveness of the attack. The strongest privacy setting, $\varepsilon = 1$, produces a ROC AUC of 0.5006, which is almost indistinguishable from random guessing. For $\varepsilon = 3$ and $\varepsilon = 10$, the AUC values increase slightly to 0.5132 and 0.5193, respectively, but remain substantially below the baseline value. This trend is consistent with the expected effect of Differential Privacy: stronger privacy guarantees introduce greater perturbation during training and consequently reduce the membership-related information available to the attacker.

The results at fixed false positive rates further illustrate the difficulty of obtaining reliable membership predictions. At an FPR of 1%, the TPR remains extremely low for all configurations, ranging from 0.88% for $\varepsilon = 3$ to 1.54% for $\varepsilon = 1$. Even when the allowed FPR is increased to 10%, the TPR remains close to 10%, with values between 8.06% and 11.16%. Therefore, the attack provides little practical discrimination in operating regions where the number of false positives must be constrained. However, the TPR values do not decrease monotonically as ε is reduced, indicating that differences between the DP configurations at individual operating points are relatively small and may be influenced

by the specific score distributions produced by each trained model.

Average Precision remains relatively stable across all experiments, ranging from 0.7990 to 0.8159. Consequently, the clearest evidence of the privacy benefit introduced by DP is provided by the ROC AUC, which moves progressively closer to random-guessing performance under the privacy-preserving configurations, particularly for $\varepsilon = 1$.

Overall, these results indicate that the considered federated models are already relatively resistant to the Quantile Membership Inference Attack, while Differential Privacy provides an additional reduction in membership leakage. The strongest privacy configuration, $\varepsilon = 1$, achieves an ROC AUC almost exactly equal to random guessing, demonstrating the greatest resistance to the attack among the evaluated settings.

4.6.2 Backdoor Attack

To evaluate the robustness of the federated learning framework against model poisoning, a Frequency-Injection Backdoor Attack (FIBA) was performed. During training, one of the three participating clients acted maliciously by injecting a frequency-domain trigger into a subset of its local training samples and assigning the poisoned samples to a predefined target class. The objective of the attacker was to introduce a hidden behaviour into the global model such that inputs containing the trigger would be classified as the target label while preserving the model's predictive performance on clean data.

The effectiveness of the attack was evaluated using two complementary metrics: Benign Accuracy (BA) and Attack Success Rate (ASR). BA measures the classification accuracy of the global model on clean test samples, whereas ASR represents the proportion of triggered samples classified as the attacker's target class, as described in Table 3.8. A successful and stealthy backdoor is therefore characterized by simultaneously high BA and high ASR, since the malicious behaviour should be activated by the trigger without substantially affecting the normal behaviour of the model.

Table 4.10 summarizes the results obtained for the four evaluated configurations.

Configuration	**DP**	**Benign Accuracy (%)**	**ASR (%)**
FedAvg	No	88.63	97.51
FedAvg	Yes	50.89	70.41
CKADA	No	16.92	100.00
CKADA	Yes	16.92	100.00

Table 4.10: Backdoor attack performance under different aggregation and privacy configurations.

Discussion

The FedAvg configuration without Differential Privacy exhibits the clearest evidence of a successful backdoor attack. The global model maintains a Benign Accuracy of approximately 88.6%, while the ASR reaches approximately 97.5%. This combination indicates that the malicious client successfully introduced the trigger behaviour without substantially compromising the predictive performance of the global model on benign samples.

Standard FedAvg is therefore highly vulnerable to the implemented attack, as the malicious update is directly incorporated into the weighted aggregation process.

The introduction of Differential Privacy substantially modifies this behaviour. Under FedAvg with DP, the ASR decreases to approximately 70.4%, indicating that gradient clipping and noise injection reduce the influence of the malicious signal contained in the poisoned client's update. However, Benign Accuracy also decreases to approximately 50.9%. The reduction in attack effectiveness is therefore accompanied by a considerable loss in model utility. Furthermore, the remaining ASR shows that Differential Privacy does not completely prevent the backdoor from being learned, which is consistent with its primary objective of limiting information leakage rather than explicitly detecting malicious client updates.

A different behaviour is observed for CKADA. Both the standard CKADA configuration and CKADA combined with Differential Privacy reach an ASR of 100%, while Benign Accuracy remains close to 16.9%. In this case, the high ASR should not be interpreted as evidence of a successful stealthy backdoor, since the global model no longer preserves acceptable performance on benign samples. Instead, the severe degradation in Benign Accuracy indicates that the training process has been strongly disrupted.

Although CKADA does not provide effective robustness against the implemented FIBA attack, this behaviour may still be considered beneficial from a monitoring perspective. Unlike FedAvg, where the backdoor remains largely hidden while benign performance stays high, the CKADA configurations exhibit an evident collapse in model utility. Such a degradation can act as a warning signal that the aggregation process has been affected by anomalous or malicious client updates. Therefore, CKADA does not prevent the attack in this experiment, but it reduces its stealthiness by making the compromised behaviour easier to identify through standard performance monitoring.

4.6.3 Gradient Inversion Attack

To evaluate the vulnerability of the proposed federated learning framework to information leakage through model updates, a Gradient Inversion Attack (GIA) was performed using the InverseFed reconstruction framework. The attack attempts to reconstruct the original training images from the model updates exchanged during federated learning, thereby assessing the extent to which sensitive patient information can be recovered by an adversary.

Unlike the Membership Inference Attack, which targets the converged global model, the Gradient Inversion Attack was applied directly to the model updates generated by an individual client after local training. The attack was conducted under the baseline federated learning configuration and repeated using model updates protected with different Homomorphic Encryption ratios. This experimental setup enables the influence of selective Homomorphic Encryption on the amount of information exposed through client updates to be evaluated.

The reconstructed images were subsequently assessed both qualitatively, through visual inspection, and quantitatively using the reconstruction quality metrics described in Table 3.8. By progressively increasing the proportion of encrypted model parameters,

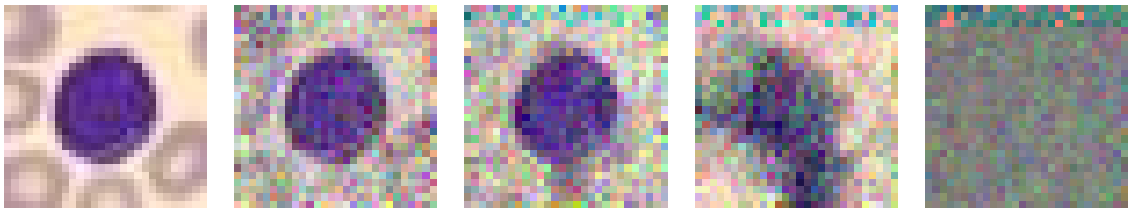
the experiments investigate whether selective Homomorphic Encryption reduces the information available to the attacker while maintaining the utility of the federated learning framework.

Table 4.11 reports the quantitative reconstruction metrics obtained for the evaluated encryption ratios.

Table 4.11: Gradient inversion attack performance under different Homomorphic Encryption ratios.

HE Ratio	MSE↓	PSNR↑	MS-SSIM↑	UQI↑	VIF↑	FMSE↓
Baseline	0.0831	16.824	0.6737	0.7165	0.5546	0.0007
25%	0.0944	16.27	0.5755	0.5933	0.4139	0.0065
50%	0.1956	13.1058	0.3717	0.2829	0.1974	0.1205
100%	0.372	10.3146	0.1968	0.0174	0.0284	3.2363

Representative reconstruction results obtained for the different encryption ratios are shown in Figure 4.4.



(a) Original (b) 0% (c) 25% (d) 50% (e) 100%

Figure 4.4: Original BloodMNIST image and its reconstructions obtained using Gradient Inversion under different Homomorphic Encryption ratios.

Discussion

The quantitative results reported in Table 4.11 demonstrate that increasing the proportion of encrypted model parameters progressively reduces the effectiveness of the Gradient Inversion Attack. The baseline configuration, in which no encryption is applied, produces the highest-quality reconstruction, achieving the lowest MSE (0.0831), the highest PSNR (16.82 dB), and the best perceptual similarity metrics, including MS-SSIM, UQI, and VIF. These results indicate that unprotected model updates contain sufficient information to reconstruct a close approximation of the original training image.

Applying selective Homomorphic Encryption noticeably degrades the reconstruction quality. With an encryption ratio of 25%, only a slight reduction in image fidelity is observed, suggesting that encrypting a limited subset of the model parameters provides only partial protection against information leakage. As the encryption ratio increases to 50%, the reconstruction quality deteriorates substantially, with the MSE more than doubling and all perceptual similarity metrics decreasing considerably. The reconstructed image also becomes visibly less similar to the original sample.

The strongest protection is achieved when all model parameters are encrypted. Under the 100% encryption configuration, the reconstruction quality degrades significantly, with the highest MSE (0.3720), the lowest PSNR (10.31 dB), and near-zero values for UQI and VIF. As illustrated in Figure 4.4, the reconstructed image contains very little visually meaningful information, indicating that the attack is no longer able to recover the original sample with acceptable fidelity.

Overall, the experimental results demonstrate that increasing the encryption ratio substantially reduces the information that can be extracted from client updates by a Gradient Inversion Attack. Although selective Homomorphic Encryption introduces additional communication and runtime overhead, as discussed in Section 4.5, it provides progressively stronger protection against information leakage. These findings suggest that selectively encrypting sensitive model parameters represents an effective compromise between privacy preservation and system efficiency, while full encryption offers the highest level of resistance to reconstruction attacks.

4.7 General Discussion

The experiments presented in this chapter provide an integrated view of the main trade-offs involved in designing a privacy-preserving and secure federated learning framework for healthcare applications. The results show that no single mechanism simultaneously maximizes predictive performance, privacy protection, security robustness, and computational efficiency. Instead, each protection mechanism introduces specific benefits and costs that must be considered jointly when selecting an appropriate system configuration.

4.7.1 Privacy-Utility Trade-off

Differential Privacy provides the clearest example of the trade-off between privacy protection and model utility. The experiments show that stronger privacy guarantees reduce the amount of information that can be extracted from the trained model and its updates. In particular, the Membership Inference Attack becomes increasingly ineffective as the privacy budget is reduced, with the strongest privacy configuration producing performance close to random guessing.

However, these privacy benefits are accompanied by a degradation in predictive performance. Stronger privacy settings introduce larger perturbations during local optimization, which affect the model's ability to learn from the original training signal. Consequently, reducing the privacy budget ε improves resistance to privacy attacks but may also decrease classification performance and increase the number of communication rounds required to reach convergence.

Differential Privacy also introduces a non-negligible computational cost, as the additional privacy-preserving operations increase the execution time of each communication round. Therefore, the selection of the privacy budget should not be based exclusively on the level of privacy protection, but should also consider its impact on model utility and training efficiency.

Homomorphic Encryption exhibits a different privacy–utility relationship. Because HE protects model parameters during transmission and aggregation without intentionally

perturbing their numerical values, predictive performance remains broadly comparable to the corresponding plaintext configurations. Its main cost is instead associated with communication and cryptographic processing. The experimental results show that selective encryption introduces substantially lower communication overhead than full-model encryption, highlighting its potential as a more efficient approach for protecting sensitive model components while limiting the additional system cost.

Overall, the experiments demonstrate that Differential Privacy and Homomorphic Encryption address different dimensions of privacy. DP protects against information inference from model updates by deliberately perturbing the learning process, while HE protects the confidentiality of the transmitted parameters during communication and aggregation. Their associated trade-offs are therefore complementary: DP primarily affects model utility and training time, whereas HE primarily affects communication and cryptographic cost.

4.7.2 Security-Performance Trade-off

The security experiments further demonstrate that improving robustness against adversarial behaviour may conflict with preserving model performance. This is particularly evident in the backdoor experiments, where the relationship between attack success and benign model behaviour must be considered jointly.

Under standard FedAvg, the FIBA backdoor attack achieves a high Attack Success Rate while maintaining strong Benign Accuracy. This represents the most concerning scenario because the malicious behaviour remains largely hidden during normal model evaluation. Differential Privacy reduces the effectiveness of the attack under FedAvg, but this reduction is accompanied by a considerable degradation in benign accuracy. DP therefore provides partial mitigation against the malicious update, although it should not be interpreted as a dedicated defence against model poisoning.

The CKADA configurations exhibit a different behaviour. Although the attack reaches a very high ASR, benign model performance collapses. Consequently, CKADA does not successfully prevent the implemented attack, but the compromised training behaviour becomes considerably less stealthy. From an operational perspective, such a degradation may be useful because standard performance monitoring could reveal that the global training process has been affected by anomalous or malicious updates. CKADA should therefore not be considered robust against the evaluated FIBA attack, but its behaviour may make severe poisoning easier to identify than in the FedAvg configuration, where the backdoor remains hidden behind high benign performance.

The privacy attacks show a different relationship between security and performance. Differential Privacy substantially improves resistance to Membership Inference, but the increased privacy protection comes at the cost of reduced predictive utility and longer training times. Homomorphic Encryption, in contrast, protects parameters while they are exchanged and aggregated without substantially modifying model predictions, but does not independently prevent attacks that exploit the behaviour of the final trained model.

These findings emphasize that the evaluated mechanisms address different threat categories. Robust aggregation is intended to limit the influence of anomalous client updates,

Differential Privacy limits information leakage from individual training samples and updates, and Homomorphic Encryption protects parameters during communication and aggregation. Consequently, combining multiple mechanisms can provide broader protection than relying on any individual defence, but it also increases system complexity and resource requirements.

4.7.3 Practical Implications for Healthcare

The results have important implications for the deployment of federated learning in healthcare environments. Medical institutions typically operate under strict privacy requirements while also facing practical limitations in computational resources, network capacity, and training time. A privacy-preserving federated learning system must therefore provide sufficient protection without making collaborative training prohibitively expensive or significantly degrading diagnostic performance.

Differential Privacy can provide meaningful protection against attacks that attempt to infer information about individual training samples. This property is particularly relevant in healthcare, where model updates may indirectly contain information derived from sensitive patient data. However, the observed reduction in model utility under stronger privacy settings indicates that highly restrictive privacy budgets may not always be appropriate when predictive accuracy is critical. The privacy parameter should therefore be selected according to the sensitivity of the application and the acceptable loss in model performance.

Homomorphic Encryption can complement DP by protecting model parameters while they are transmitted and aggregated. Nevertheless, the large communication overhead observed under full encryption may limit its practicality in environments with constrained network resources. The selective encryption strategy evaluated in this work provides a more practical alternative by encrypting only a subset of the model parameters. This substantially reduces communication requirements while preserving the main advantage of encrypted aggregation for the selected components.

The backdoor experiments additionally highlight the importance of continuous monitoring during federated training. High predictive performance alone cannot guarantee that the global model has not been compromised, as demonstrated by the FedAvg configuration in which high benign accuracy coexists with a high attack success rate. Conversely, a sudden deterioration in benign performance may provide an indication that the aggregation process is being affected by abnormal client behaviour. Practical deployments should therefore monitor not only global accuracy but also client updates, convergence behaviour, class-level predictions, and other indicators that may reveal anomalous training dynamics.

Overall, the experimental results suggest that a layered protection strategy is more appropriate for healthcare federated learning than relying on a single privacy or security mechanism. Differential Privacy can reduce information leakage, Homomorphic Encryption can protect model updates during communication, and robust aggregation can provide an additional mechanism for managing potentially anomalous client contributions. However, these protections should be configured selectively according to the application requirements. In particular, moderate privacy budgets and selective encryption may offer

a more practical balance between privacy, security, predictive performance, communication efficiency, and computational cost than applying the strongest possible protection to every component of the federated learning process.

Chapter 5

Conclusion

5.1 Summary of Findings

This thesis investigated the application of privacy-preserving Federated Learning for healthcare image classification. The proposed framework enables multiple healthcare institutions to collaboratively train a deep learning model without exchanging their raw patient data, thereby addressing one of the principal barriers to large-scale collaborative artificial intelligence in healthcare.

A complete experimental evaluation was conducted using the BloodMNIST dataset under both centralized and federated learning settings. The study investigated the influence of different data distributions, aggregation strategies, and privacy-preserving mechanisms on predictive performance, computational efficiency, and robustness against representative privacy and security attacks.

The experimental results demonstrate that Federated Learning can achieve predictive performance comparable to centralized learning while preserving data locality. Furthermore, the evaluation highlights the trade-offs introduced by Differential Privacy and Homomorphic Encryption, showing that stronger privacy protection generally increases computational overhead and may reduce model utility. Security experiments further demonstrate the effectiveness of these mechanisms in mitigating Membership Inference, Gradient Inversion, and Backdoor attacks.

5.2 Research Contributions

This thesis contributes to the study of privacy-preserving Federated Learning for healthcare by designing and implementing a complete federated learning framework for medical image classification using the Flower framework. The proposed architecture integrates multiple privacy and security mechanisms within the same experimental environment, including Differential Privacy and selective Homomorphic Encryption.

Different aggregation strategies were implemented and evaluated, including FedAvg, CKADA, and a hybrid aggregation mechanism combining robust aggregation with encrypted parameter aggregation. The framework was also assessed under both IID and non-IID data distributions in order to investigate the influence of statistical heterogeneity

on federated model performance.

Beyond predictive performance, the thesis provides a computational and communication analysis of the evaluated configurations, considering execution time, communication overhead, and floating-point operations (FLOPs). This allows the practical cost introduced by federated training and privacy-preserving mechanisms to be examined alongside their benefits.

The security of the framework was further investigated against representative threats affecting federated learning, including Membership Inference, Gradient Inversion, and Frequency-Injection Backdoor attacks. These experiments provide insight into the extent to which the considered privacy and aggregation mechanisms can reduce information leakage or mitigate malicious behaviour.

Overall, the main contribution of this work is an integrated experimental analysis of the trade-offs between predictive performance, privacy preservation, security, and computational efficiency in healthcare Federated Learning. Rather than evaluating these aspects independently, the proposed framework enables their interactions to be examined within a common experimental setting.

5.3 Limitations

Although the proposed framework demonstrates promising results, several limitations should be acknowledged.

First, all experiments were conducted using the BloodMNIST dataset, which represents a relatively small medical imaging benchmark. Although suitable for controlled experimentation, larger and more heterogeneous clinical datasets would provide a more comprehensive evaluation of the proposed framework.

Second, the federated learning environment was simulated using a limited number of participating clients. While this configuration enables reproducible experimentation, real-world healthcare deployments may involve substantially larger numbers of institutions with more diverse computational resources, communication networks, and data distributions.

Third, the security evaluation focused on three representative attack classes: Membership Inference, Gradient Inversion, and Backdoor attacks. Other threats, including model extraction, Byzantine attacks, and poisoning strategies targeting the aggregation process, were outside the scope of this work.

Finally, only selected privacy-preserving mechanisms were investigated. Alternative approaches, such as Secure Multi-Party Computation, Trusted Execution Environments, or personalized privacy mechanisms, may provide different trade-offs between privacy protection and computational efficiency.

5.4 Future Research Directions

The work presented in this thesis provides a foundation for privacy-preserving Federated Learning in healthcare; however, several aspects require further investigation to strengthen the applicability and generalizability of the proposed framework.

First, future work should evaluate the proposed framework on larger and more diverse medical datasets involving multiple imaging modalities and clinical data sources. The experiments presented in this thesis were conducted using the BloodMNIST dataset, which provides a suitable benchmark for controlled experimentation but cannot fully represent the complexity and variability of real-world healthcare data. Assessing the framework on multi-institutional clinical datasets would provide stronger evidence of its generalizability.

A second direction concerns scalability. The experimental evaluation considered a simulated cross-silo federated learning environment with a limited number of participating clients. Future studies should investigate larger federated networks involving a greater number of healthcare institutions with heterogeneous computational resources, communication infrastructures, and data distributions. Such experiments would provide a more realistic assessment of the scalability and robustness of the proposed framework.

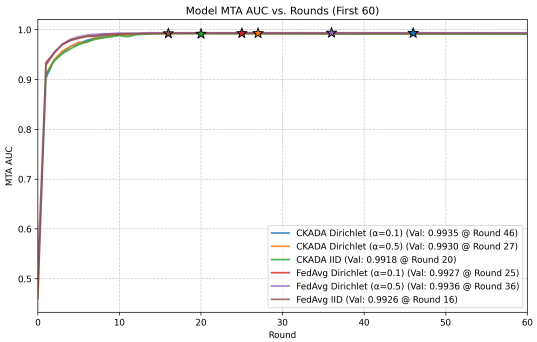
Further research should also extend the security evaluation by considering additional attack scenarios. Although this thesis investigated Membership Inference, Gradient Inversion, and Backdoor attacks, other threats, including model extraction, Byzantine attacks, data poisoning strategies, and adaptive adversarial attacks, should be evaluated to provide a more comprehensive security assessment.

Another promising direction is the investigation of additional privacy-preserving mechanisms and adaptive defense strategies. Future work may compare the proposed framework with Secure Multi-Party Computation, Trusted Execution Environments, or personalized Differential Privacy schemes, as well as develop adaptive approaches that dynamically balance privacy protection, computational cost, and model utility according to the characteristics of individual clients or model layers.

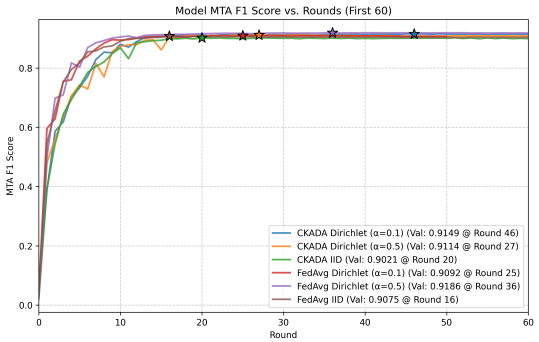
Finally, future research should further investigate the practical deployment of Federated Learning within real healthcare environments. This includes validating the proposed framework under realistic communication conditions, integrating it with existing hospital information systems, and assessing its compliance with emerging regulatory frameworks such as the European AI Act. These studies would strengthen the external validity of the proposed approach and provide additional evidence regarding its applicability in real-world clinical settings.

Appendix A

Additional Experimental Results

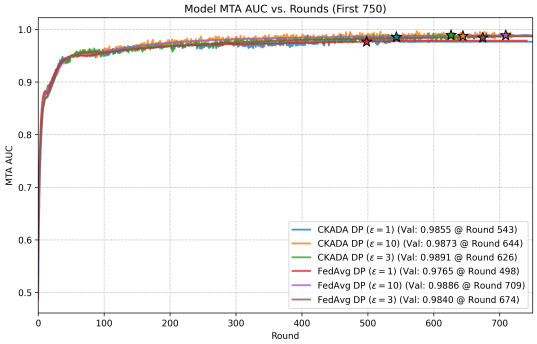


(a) Evolution of the AUC during federated training.

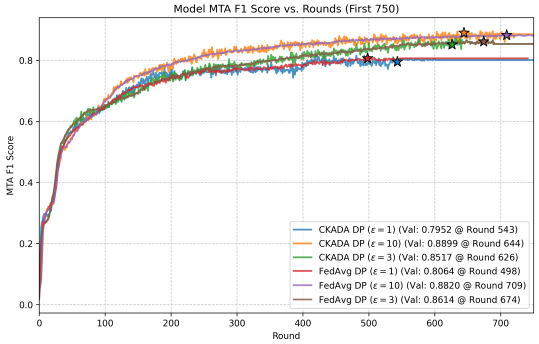


(b) Evolution of the F1-score during federated training.

Figure A.1: Evolution of the AUC and F1-score during federated training.



(a) Evolution of the AUC during federated training.



(b) Evolution of the F1-score during federated training.

Figure A.2: Evolution of the AUC and F1-score during federated training with DP enabled.

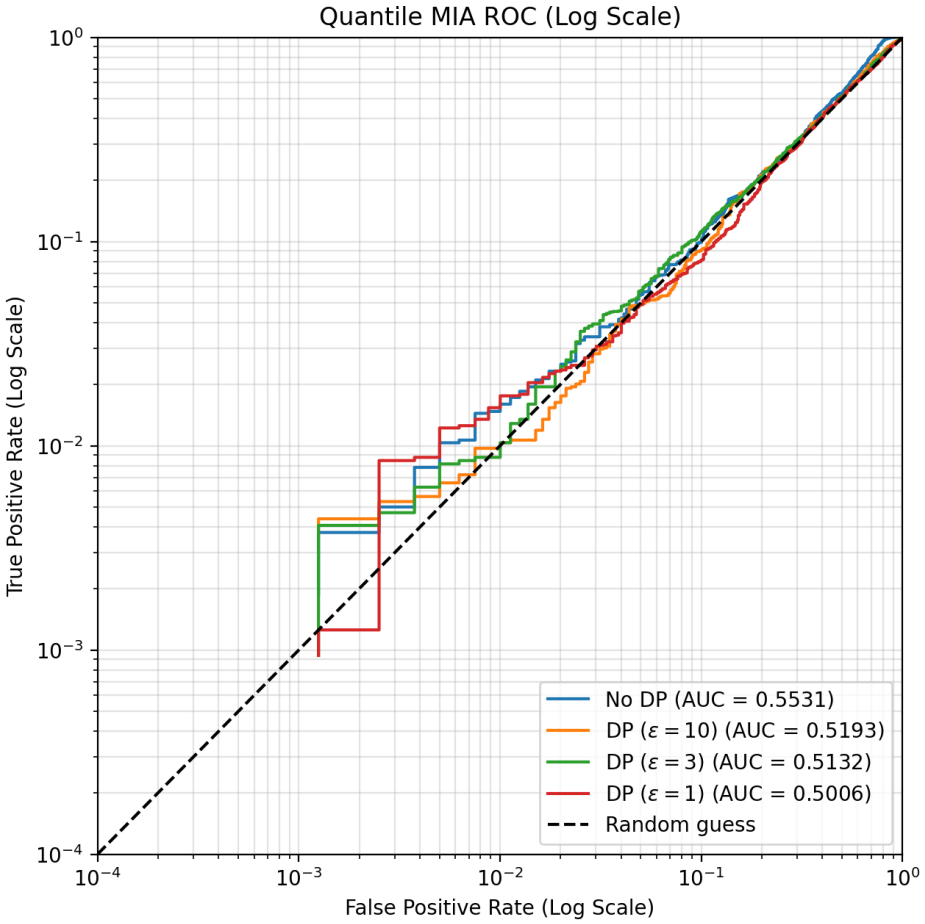


Figure A.3: ROC curves of the Quantile Membership Inference Attack under different Differential Privacy configurations, shown on a logarithmic scale.

Bibliography

[1] Aysu Acar, Hidayet Aksu, Selcuk Uluagac, and Mauro Conti. A survey on homomorphic encryption schemes: Theory and implementation. *ACM Computing Surveys*, 51(4):1–35, 2018.

[2] Jane Andrew and Max Baker. The general data protection regulation in the age of surveillance capitalism: J. andrew, m. baker. *Journal of Business Ethics*, 168(3):565–578, 2021.

[3] Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 2938–2948, 2020.

[4] Kornelia Batko and Andrzej Ślęzak. The use of big data analytics in healthcare. *Journal of big Data*, 9(1):3, 2022.

[5] Ayoub Benaissa, Bilal Retiat, Bogdan Cebere, and Alaa Eddine Belfedhal. TenSEAL: A library for encrypted tensor operations using homomorphic encryption. *arXiv preprint arXiv:2104.03152*, 2021.

[6] Daniel J. Beutel, Taner Topal, Akhil Mathur, Xinchi Qiu, Titouan Parcollet, and Nicholas D. Lane. Flower: A friendly federated learning research framework. *CoRR*, abs/2007.14390, 2020.

[7] Peva Blanchard et al. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Advances in Neural Information Processing Systems*, 2017.

[8] Nader Bouacida and Prasant Mohapatra. Vulnerabilities in federated learning. *IEEE Access*, 9:63229–63249, 2021.

[9] Zhiqi Bu, Jialin Mao, and Shiyun Xu. Scalable and efficient training of large convolutional neural networks with differential privacy, 2022.

[10] Ayushi Chahal and Preeti Gulia. Machine learning and deep learning. *International Journal of Innovative Technology and Exploring Engineering*, 8(12):4910–4914, 2019.

[11] Jung Hee Cheon, Andrey Kim, Miran Kim, and Yongsoo Song. Homomorphic encryption for arithmetic of approximate numbers. In Tsuyoshi Takagi and Thomas Peyrin, editors, *Advances in Cryptology – ASIACRYPT 2017*, volume 10624 of *Lecture Notes in Computer Science*, pages 409–437. Springer, 2017.

[12] Rafail C Christodoulou, Amanda Woodward, Rafael Pitsillos, Reina Ibrahim, and Michalis F Georgiou. Artificial intelligence in alzheimer's disease diagnosis and prognosis using pet-mri: A narrative review of high-impact literature post-tauvid approval. *Journal of Clinical Medicine*, 14(16):5913, 2025.

[13] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pages

265–284. Springer, 2006.

[14] El Mahdi El Mhamdi, Rachid Guerraoui, and Sebastien Rouault. The hidden vulnerability of distributed learning in byzantium. In *Proceedings of the 35th International Conference on Machine Learning*, pages 3521–3530, 2018.

[15] Yosri A. Fahim, Ibrahim W. Hasani, Samer Kabba, and Waleed Mahmoud Ragab. Artificial intelligence in healthcare and medicine: clinical applications, therapeutic advances, and future perspectives. *European Journal of Medical Research*, 30(848), 2025.

[16] Yu Feng, Benteng Ma, Jing Zhang, Shanshan Zhao, Yong Xia, and Dacheng Tao. Fiba: Frequency-injection based backdoor attack in medical image analysis. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20844–20853. IEEE, 2022.

[17] Wensheng Gan, Kaixia Hu, Gengsen Huang, Wei-Che Chien, Han-Chieh Chao, and Weizhi Meng. Data analytic for healthcare cyber physical system. *IEEE Transactions on Network Science and Engineering*, 10(5):2490–2502, 2023.

[18] Teja Reddy Gatla. An innovative study exploring revolutionizing healthcare with ai: Personalized medicine: Predictive diagnostic techniques and individualized treatment. *International Journal of Advanced Research and Interdisciplinary Scientific Endeavours*, 1(2):61–70, 2024.

[19] Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients-how easy is it to break privacy in federated learning? *Advances in neural information processing systems*, 33:16937–16947, 2020.

[20] Craig Gentry. A fully homomorphic encryption scheme. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing*, pages 169–178, 2009.

[21] Robin Geyer, Tassilo Klein, and Moin Nabi. Differentially private federated learning: A client level perspective. *arXiv preprint arXiv:1712.07557*, 2017.

[22] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. In *arXiv preprint arXiv:1708.06733*, 2017.

[23] Hao Guan, Pew-Thian Yap, Andrea Bozoki, and Mingxia Liu. Federated learning for medical image analysis: A survey. *Pattern Recognition*, 151:110424, 2024.

[24] Chaoyang He, Songze Li, Jinhyun So, Mi Zhang, Hongyi Wang, Xiaoyang Wang, Praneeth Vepakomma, Abhishek Singh, Hang Qiu, Li Shen, Peilin Zhao, Yan Kang, Yang Liu, Ramesh Raskar, Qiang Yang, Murali Annavaram, and Salman Avestimehr. Fedml: A research library and benchmark for federated machine learning. *CoRR*, abs/2007.13518, 2020.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[26] Chris Jay Hoofnagle, Bart Van Der Sloot, and Frederik Zuiderveen Borgesius. The european union general data protection regulation: what it is and what it means. *Information & communications technology law*, 28(1):65–98, 2019.

[27] Alexander Horst, Paul Loustalot, Sanjeev Yoganathan, Ting Li, Joshua Xu, Weida

Tong, David Schneider, Nicolas Löffler-Perez, Erminio Di Renzo, and Michael Renaudin. Federated learning: a privacy-preserving approach to data-centric regulatory cooperation. *Frontiers in Drug Safety and Regulation*, 5:1579922, 2025.

[28] Yangsibo Huang, Samyak Gupta, Zhao Song, Kai Li, and Sanjeev Arora. Evaluating gradient inversion attacks and defenses in federated learning. *Advances in neural information processing systems*, 34:7232–7241, 2021.

[29] Agbotiname Lucky Imoize, Mohammad S. Obaidat, and Houbing Herbert Song. Chapter 14 - legal implications of federated learning integration in digital healthcare systems. In Agbotiname Lucky Imoize, Mohammad S. Obaidat, and Houbing Herbert Song, editors, *Federated Learning for Digital Healthcare Systems*, Intelligent Data-Centric Systems, pages 355–385. Academic Press, 2024.

[30] Peter Kairouz, H. Brendan McMahan, et al. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2):1–210, 2021.

[31] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. A survey on federated learning systems: Vision, hype and reality for data privacy and protection. *IEEE Transactions on Knowledge and Data Engineering*, 2020. Early Access.

[32] Ji Liu, Jizhou Huang, Yang Zhou, Xuhong Li, Shilei Ji, Haoyi Xiong, and Dejing Dou. From distributed machine learning to federated learning: A survey. *CoRR*, abs/2104.14362, 2021.

[33] Xiaofei Ma et al. Privacy-preserving federated learning based on homomorphic encryption: A survey. *Electronics*, 11(17), 2022.

[34] Priyanka Mary Mammen. Federated learning: Opportunities and challenges, 2021.

[35] Laurie T Martin, Christopher Nelson, Douglas Yeung, Joie D Acosta, Nabeel Qureshi, Tara Blagg, and Anita Chandra. The issues of interoperability and data connectedness for public health. *Big data*, 10(1_suppl):S19–S24, 2022.

[36] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR, 20–22 Apr 2017.

[37] H. Brendan McMahan et al. Learning differentially private recurrent language models. *International Conference on Learning Representations*, 2018.

[38] Veera Mothukuri, Reza M. Parizi, Seyedamin Pouriyeh, Yuhong Huang, Ali Dehghantanha, and Gautam Srivastava. A survey on security and privacy of federated learning. *Future Generation Computer Systems*, 115:619–640, 2021.

[39] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[40] Anichur Rahman, Md Sazzad Hossain, Ghulam Muhammad, Dipanjali Kundu, Tanoy Debnath, Muaz Rahman, Md Saikat Islam Khan, Prayag Tiwari, and Shahab S Band. Federated learning-based ai approaches in smart healthcare: concepts, taxonomies, challenges and open issues. *Cluster computing*, 26(4):2271–2311, 2023.

[41] Nuria Rodríguez-Barroso, Daniel Jiménez-López, M Victoria Luzón, Francisco Herrera, and Eugenio Martínez-Cámara. Survey on federated learning threats: Concepts, taxonomy on attacks and defences, experimental study and challenges. *Information Fusion*, 90:148–173, 2023.

[42] Sinem Sav et al. Poseidon: Privacy-preserving federated neural network learning. In *Network and Distributed System Security Symposium*, 2019.

[43] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy*, pages 3–18. IEEE, 2017.

[44] Stacey Truex, Ling Liu, Mehmet Emre Gursoy, Lei Yu, and Wenqi Wei. Ldp-fed: Federated learning with local differential privacy. *Proceedings of the Third ACM International Workshop on Edge Systems, Analytics and Networking*, pages 61–66, 2020.

[45] Nguyen Truong, Kai Sun, Siyao Wang, Florian Guitton, and Yike Guo. Privacy preservation in federated learning: An insightful survey from the gdpr perspective, 2021.

[46] Xiaofang Wang, Qianfang Jia, Lvyuan Liang, Weiwei Zhou, Weihua Yang, and Jingfeng Mu. Artificial intelligence in adhd: a global perspective on research hotspots, trends and clinical applications. *Frontiers in Human Neuroscience*, 19:1577585, 2025.

[47] Lauren Watson, Chuan Guo, et al. Scalable membership inference attacks via quantile regression. In *International Conference on Learning Representations*, 2022.

[48] Rolf H. Weber. Legal interoperability as a tool for combatting fragmentation, December 2014.

[49] Kang Wei et al. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Transactions on Information Forensics and Security*, 2020.

[50] Jie Wen, Zhixia Zhang, Yang Lan, Zhihua Cui, Jianghui Cai, and Wensheng Zhang. A survey on federated learning: challenges and applications. *International journal of machine learning and cybernetics*, 14(2):513–535, 2023.

[51] Herbert WoisetschlÃ¤ger, Alexander Erben, Bill Marino, Shiqiang Wang, Nicholas D. Lane, Ruben Mayer, and Hans-Arno Jacobsen. Federated learning priorities under the european union artificial intelligence act, 2024.

[52] Chulin Xie, Keli Huang, Pin-Yu Chen, and Bo Li. Dba: Distributed backdoor attacks against federated learning. In *International Conference on Learning Representations*, 2020.

[53] Jiancheng Yang, Rui Shi, and Bingbing Ni. MedMNIST classification decathlon: A lightweight AutoML benchmark for medical image analysis. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 191–195. IEEE, 2021.

[54] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. MedMNIST v2: A large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

[55] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *IEEE Computer Security Foundations Symposium*, pages 268–282. IEEE, 2018.

[56] Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *Proceedings of the 35th International Conference on Machine Learning*, pages 5650–5659, 2018.

[57] Hongxu Yin, Arun Mallya, Arash Vahdat, Jose M Alvarez, Jan Kautz, and Pavlo Molchanov. See through gradients: Image batch recovery via gradinversion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16337–16346, 2021.

[58] Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, Graham Cormode, and Ilya Mironov. Opacus: User-friendly differential privacy library in PyTorch. *arXiv preprint arXiv:2109.12298*, 2021.

[59] Bocong Yuan and Jiannan Li. The policy effect of the general data protection regulation (gdpr) on the digital public health sector in the european union: an empirical investigation. *International journal of environmental research and public health*, 16(6):1070, 2019.

[60] Qiang Zhang, Xiaoli Chen, Shiwei Zhong, et al. A survey on privacy-preserving federated learning. *Knowledge-Based Systems*, 250:109092, 2022.

[61] Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. *Advances in neural information processing systems*, 32, 2019.

[62] Tengteng Zhu, Zehua Guo, Chao Yao, Jiaxin Tan, Songshi Dou, Wenrun Wang, and Zhenzhen Han. Byzantine-robust federated learning via cosine similarity aggregation. *Computer Networks*, 254:110730, 2024.