

# A Secured Federated Learning System Leveraging Confidence Score to Identify Retinal Disease

by

M Sakib Osman Eshan  
19101412

Md. Naimul Huda Nafi  
19101400

Nazmus Sakib  
19101404

Md. Ahnaf Morshed Maruf  
20101630

Mehedi Hasan Emon  
19301234

A thesis submitted to the Department of Computer Science and Engineering in partial fulfillment of the requirements for the degree of B.Sc. in Computer Science

Department of Computer Science and Engineering  
School of Data and Sciences  
Brac University  
May 2023

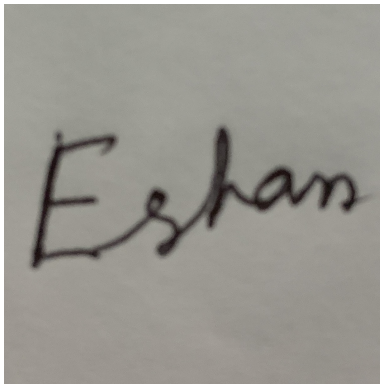
© 2023. Brac University  
All rights reserved.

# Declaration

It is hereby declared that

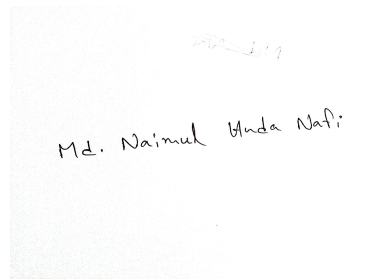
1. The final thesis submitted is my/our own original work while completing degree at BRAC University.
2. The final thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The final thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

**Student's Full Name & Signature:**



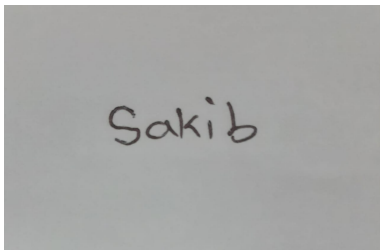
---

M Sakib Osman Eshan  
19101412



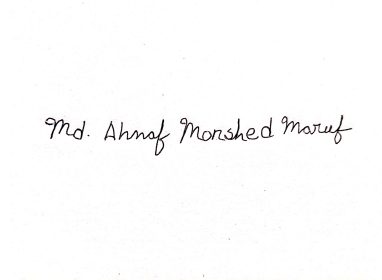
---

Md. Naimul Huda Nafi  
19101400



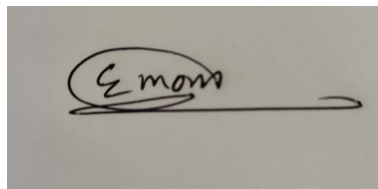
---

Nazmus Sakib  
19101404



---

Md. Ahnaf Morshed Maruf  
20101630



---

Mehedi Hasan Emon  
19301234

# Approval

The thesis/project titled “A Secured Federated Learning System Leveraging Confidence Score to Identify Retinal Disease” submitted by

1. M Sakib Osman Eshan(19101412)
2. Md. Naimul Huda Nafi(19101400)
3. Nazmus Sakib(19101404)
4. Md. Ahnaf Morshed Maruf(20101630)
5. Mehedi Hasan Emon(19301234)

Of spring, 2023 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on May 22, 2023.

## Examining Committee:

Supervisor 1:

*Rahman*

---

Rafeed Rahman  
Lecturer  
Department of Computer Science and Engineering  
Brac University

Supervisor 2:

*Tanzim*

---

Tanzim Reza  
Lecturer  
Department of Computer Science and Engineering  
Brac University

Co-supervisor:



---

Dr. Mohammad Zavid Parvez Sir  
Academic  
School of Computing, Mathematics and Engineering  
Charles Sturt University, Australia

Chairperson:

---

Sadia Hamid Kazi  
Chairperson and Associate Professor  
Department of Computer Science and Engineering  
Brac University

Coordinator:

---

Md. Golam Rabiul Alam  
Professor  
Department of Computer Science and Engineering  
Brac University

# Abstract

Federated learning is a distributed machine learning paradigm that enables multiple clients to collaboratively train a global model without sharing their local data. However, federated learning is vulnerable to adversarial attacks, where malicious clients can manipulate their local updates to degrade the performance or compromise the privacy of the global model. To mitigate this problem, this paper proposes a novel method that reduces the influence of malicious clients based on their confidence. We conducted our experiments on the Retinal OCT dataset. The proposed technique significantly improves the global model's precision, recall, F1 score, and area under the receiver operating characteristic curve (AUC-ROC). Precision rises from 0.869 to 0.906, recall rises from 0.836 to 0.889, F1 score rises from 0.852 to 0.898, and AUC-ROC rises from 0.836 to 0.889.

**Keywords:** Computer Vision; Federated learning; Deep Learning; Healthcare; Data poisoning; Retinal OCT

# Dedication

By dedicating this thesis paper to our parents, who have been our rock throughout this entire process, we wish to express our sincerest gratitude. Thank you, Mom and Dad, for loving us, supporting us, and making so many sacrifices so that we could become the individuals we are today. Your unwavering trust in us through thick and lean has provided us with the fortitude to persevere despite these academic challenges.

You have been our guides, providing us with advice and information that has increased our comprehension and broadened our horizons. Without your unending encouragement and altruism, it would have been impossible for us to achieve our objectives.

Your commitment to teaching us the importance of education and the pleasure of learning has been priceless. Your unwavering faith in us has kept us going despite the arduous labor, frustration, and uncertainty we've encountered along the way.

This would not have been feasible without your unwavering support, altruistic contributions, and numerous cheers. We want you to know how much we value your assistance and how much this thesis report means to us.

With profound affection and appreciation,

M Sakib Osman Eshan  
Md. Naimul Huda Nafi  
Nazmus Sakib  
Md. Ahnaf Morshed Maruf  
Mehedi Hasan Emon  
May 2023

## **Acknowledgement**

Firstly, all praise to the Great Allah for whom our thesis have been completed without any major interruption.

Secondly, to our co-supervisor Dr. Mohammad Zavid Parvez Sir for his kind support and advice in our work. He helped us whenever we needed help.

Thirdly, to our supervisor Tanzim Reza and Rafeed Rahman Sir for their support.

And finally to our parents without their throughout support it may not be possible. With their kind support and prayer we are now on the verge of our graduation.

# Table of Contents

<b>Declaration</b>	<b>i</b>
<b>Approval</b>	<b>ii</b>
<b>Abstract</b>	<b>iv</b>
<b>Dedication</b>	<b>v</b>
<b>Acknowledgment</b>	<b>vi</b>
<b>Table of Contents</b>	<b>vii</b>
<b>List of Figures</b>	<b>ix</b>
<b>List of Tables</b>	<b>x</b>
<b>Nomenclature</b>	<b>x</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Problem Statement . . . . .	4
1.2 Research Objectives . . . . .	5
<b>2 Literature Review</b>	<b>7</b>
<b>3 Background Study</b>	<b>18</b>
3.1 Federated learning . . . . .	18
3.2 InceptionV3 . . . . .	20
3.2.1 Architecture . . . . .	20
3.3 Label Poisoning . . . . .	21
3.4 Data Set . . . . .	23
3.5 Softmax Function . . . . .	23
<b>4 Work Plan</b>	<b>25</b>
4.1 Data curation . . . . .	25
4.2 Curation of local data from total data subset . . . . .	25
4.3 Data preprocessing . . . . .	25
4.4 Apply data poisoning . . . . .	25
4.5 Simulate a regular federated learning system and a secure federated learning system . . . . .	26
4.6 Model evaluation . . . . .	26



4.7	Compare results . . . . .	26
<b>5</b>	<b>Proposed Method and Implementation</b>	<b>28</b>
5.1	Confidence Score Based Aggregation . . . . .	28
5.2	Implementation . . . . .	28
5.2.1	Data Organization . . . . .	28
5.2.2	Data Pre-processing . . . . .	29
5.2.3	InceptionV3 . . . . .	29
5.2.4	Secure Federated Learning System . . . . .	30
5.2.5	Evaluation . . . . .	30
<b>6</b>	<b>Result Analysis</b>	<b>31</b>
6.1	Confidence Scores . . . . .	31
6.2	Accuracy and Loss . . . . .	33
6.3	Confusion Matrix . . . . .	34
<b>7</b>	<b>Conclusion and Future Works</b>	<b>36</b>
7.1	Conclusion . . . . .	36
7.2	Future Works . . . . .	36
	<b>Bibliography</b>	<b>37</b>

# List of Figures

1.1	Federated Learning Layout. . . . .	1
3.1	Federated Learning Architecture for Hospital Systems . . . . .	19
3.2	Inception V3 Model Architecture Diagram . . . . .	20
3.3	Data Poisoning Architecture Diagram . . . . .	22
3.4	Retinal OCT Dataset Samples. . . . .	23
4.1	Working plan . . . . .	27
5.1	Visualization of Local datasets . . . . .	29
6.1	Confidence score rates of the Insecure FLS. . . . .	31
6.2	Confidence score rates of the Secure FLS. . . . .	32
6.3	Accuracy rates of the Insecure FLS. . . . .	33
6.4	Loss rates of the Secure FLS. . . . .	33
6.5	Confusion matrix of the Insecure FLS. . . . .	34
6.6	Confusion matrix of the Secure FLS. . . . .	35

# List of Tables

5.1	Malicious local datasets . . . . .	29
6.1	Comparison of performance metrics between Insecure FLS and secure FLS . . . . .	35

# Chapter 1

## Introduction

Recent years have seen tremendous developments in the fields of machine learning and artificial intelligence, which have revolutionized a wide range of sectors and fields. Because of extensive datasets and robust computational resources, these developments have been possible. The distributed learning paradigm of federated learning allows for collaborative model training without the need to share raw data and is one of the most significant advancements in machine learning.

Because of its potential to alleviate privacy issues brought on by centralized data collection, federated learning has received a lot of interest. Individuals can use their own data to train machine learning models, while the models' parameters are pooled in a central location. By keeping the raw data on individual devices and only exchanging model updates, this distributed system protects the confidentiality of user information. But as federated learning grows in popularity, new difficulties and security holes appear, especially in the form of adversarial attacks.

In a federated learning environment, machine learning models are especially vulnerable to adversarial attacks. Data poisoning is a sort of attack in which bad actors insert harmful or corrupted data samples into the local datasets of other actors' devices. The correctness of the shared model can be jeopardized by the introduction of bias and misinformation through the use of poisoned samples. As a result, researchers have had to shift their attention to finding ways to prevent and lessen the impact of data poisoning assaults on federated learning.



Figure 1.1: Federated Learning Layout.

The importance of this study rests in its objective to create efficient safeguards against data poisoning assaults in federated education. This research aims to improve the safety and stability of federated learning systems by illuminating the weaknesses and dynamics of such attacks. Following is a brief synopsis of the study's most important findings and contributions:

- Since machine learning models are increasingly being used in highly sensitive areas like healthcare, banking, and personal information, protecting this data is of the utmost importance. This study aids in protecting sensitive data from malicious tampering, keeping user privacy intact, and protecting the secrecy of personal information by preventing data poisoning assaults.
- Protecting the honesty of the models used in machine learning, which, if corrupted or poisoned, could produce disastrous results in safety-critical contexts. This study seeks to protect the honesty of federated machine learning models through the creation of safeguards, allowing for more trustworthy predictions across a variety of domains.
- As AI systems become more pervasive in daily life, it is essential that we build a foundation of trust in them. By thwarting data poisoning assaults, we can have more faith in AI systems, knowing that the outcomes of their calculations will be just and free of prejudice. This study helps us create AI systems that can be relied on in practical settings.
- To protect federated learning from adversarial attacks, this study contributes to the development of countermeasures. This research adds to the current body of knowledge in the field of machine learning security and helps contribute to the development of more robust and resilient defence mechanisms by introducing unique tactics and strategies to resist data poisoning attacks.
- Adversarial assaults raise ethical questions, especially when they target highly regulated industries like healthcare and finance. This study is congruent with ethical concerns about safeguarding the veracity and trustworthiness of AI systems, safeguarding the privacy of users, and guaranteeing the impartiality of decision-making by concentrating on the prevention of data poisoning assaults in federated learning.
- The consequences of this study go far beyond the academic community and into the business and social worlds. Industries, government agencies, and society at large all stand to greatly benefit from the prevention of data poisoning assaults in federated learning. It protects the well-being and interests of individuals and businesses by preserving the honesty of AI systems employed in life-or-death applications, including medical diagnosis, financial fraud detection, and cybersecurity.

A number of statistics and academic sources support the significance of this study.

Cybercrime is expected to cost the global economy 6 trillion dollar by 2021, according to a report by the World Economic Forum. This shows the increasing importance of implementing stringent security mechanisms in machine learning systems, such as federated learning, to prevent malicious assaults.

In addition, Tramèr et al.[2] demonstrated how simple it is for data poisoning attacks to trick machine learning algorithms. They demonstrated that even introducing a negligible amount of tainted data into the training phase can have a major impact on the quality of the resulting model. As a result, it is more crucial than ever to create failsafes that protect the confidentiality of federated educational networks.

Because of the sensitive nature of patient data, the repercussions of a data poisoning assault in the healthcare sector can be devastating. Misdiagnosis and unnecessary medical procedures may result from adversarial attacks on medical image classification models, according to research by Li et al.[3]. To guarantee patient safety and keep healthcare providers' faith in AI systems, it is crucial to prevent data poisoning assaults on these systems.

In addition, malicious attacks, such as credit card fraud and money laundering, pose serious threats to the financial sector. According to research by the Association for Financial Professionals[4], 82 percent of businesses will encounter either attempted or successful payment fraud in 2020. Safeguards against data poisoning attacks in federated learning can be developed to lessen their impact and boost the safety of monetary exchanges.

Although the threat that data poisoning attacks pose is becoming more widely recognised, there is still room for improvement in the design of prevention systems that are ideal for federated educational environments. However, previous research has largely ignored the particular challenges and dynamics of federated learning systems. In light of the dispersed nature of the learning process and the privacy constraints imposed by local data ownership, this study seeks to fill this void by exploring fresh methodologies and tactics to prevent data poisoning attempts in federated learning.

In this paper, we present a hypothesis that implementing a mechanism to mitigating the influence of malicious local models according to their confidence score can significantly enhance the robustness and privacy of the federated learning system.

This thesis can be broken down into the following sections: The research on adversarial assaults, data poisoning, and federated learning is reviewed extensively in Chapter 2. The methodologies and tools used in this study are detailed in Chapter 3. These tools and methods include anomaly detection algorithms, secure aggregation protocols, and adaptive learning methodologies. In Chapter 4, we detail the experimental design, data sets, and metrics utilized to test the efficacy of the proposed safety measures. Chapter 5 discusses the analysis and outcomes of the experiments. Chapter 6 concludes with a review of the study's findings, an analysis

of their significance, and suggestions for further study.

To sum up, protecting federated learning from data poisoning threats is crucial for protecting individuals' privacy and the trustworthiness of machine learning models across industries. This study helps advance secure and resilient federated learning systems, which have far-reaching ramifications for industry, society, and the ethical usage of AI, by filling in a knowledge gap and creating effective preventative mechanisms.

## 1.1 Problem Statement

In order to train models in a group setting while protecting individual privacy, federated learning has emerged as a viable option. However, it opens up hitherto unseen weaknesses, especially to malicious attacks. Data poisoning, a type of attack in which hostile users insert corrupted or malicious data samples into the local datasets of participating devices, can jeopardize the integrity and trustworthiness of machine learning models in federated learning contexts.

Multiple reports have shown that federated learning systems are susceptible to data poisoning attacks. Machine learning models were shown to be vulnerable to adversarial attacks by Tramèr et al. (2017) [5], highlighting the importance of having strong defenses. The possible impact of data poisoning assaults on the learning process was further explored by Bhagoji et al. (2018) [6] in the context of federated learning. To further emphasize the significance of addressing security concerns, Bagdasaryan et al. (2020) [7] suggested a threat model for federated learning.

In federated learning, the effects of data poisoning assaults can be devastating. Patient outcomes in healthcare, for example, could be jeopardized if a corrupted model led to inaccurate predictions (Li et al., 2019). [8]. Losses in the financial sector could rise as a result of data poisoning assaults that compromise fraud detection systems (Association of Financial Professionals, 2020) [9].

There is a knowledge gap when it comes to building effective preventive methods for this distributed learning paradigm, despite the increased awareness of data poisoning assaults in federated learning. Data poisoning threats in federated learning situations are not adequately addressed in the existing research, which is largely concerned with centralized learning settings.

This study addresses a double-sided issue. To begin, effective safeguards against data poisoning threats in federated learning must be created. These safeguards should protect the confidentiality and trustworthiness of machine learning models without compromising their security or integrity. Second, there is a void in the literature regarding how to deal with data poisoning assaults in a federated learning environment, as most studies have been conducted in a centralized learning environment.

When it comes to the first, we'll be looking into the susceptibilities and dynamics of data poisoning attacks in federated learning to see how they might be countered.

Effective preventative systems can be created by learning about assault mechanisms and how they disrupt learning. In order to detect and cope with poisoned data, we will investigate methods including anomaly detection, resilient aggregation, and model verification. The project intends to improve the safety and reliability of federated educational systems by creating these methods.

Second, this study intends to fill a vacuum in the literature by suggesting innovative preventative methods tailored to federated learning. The existing literature on data poisoning assaults in centralized learning environments lays the groundwork for knowing the fundamentals of countermeasures. However, due to federated learning's decentralized character, techniques must be adapted to take into consideration the specific difficulties and limitations of this kind of collaborative setting. By filling this void, the study hopes to promote federated learning and make it more accessible in areas where privacy is a concern.

Existing literature on topics including adversarial attacks, data poisoning, federated learning, and defense mechanisms will be thoroughly analyzed as part of the proposed study. This study will identify research gaps and the potential for building efficient preventative strategies against data poisoning assaults by synthesizing and critically analyzing the current level of knowledge in these domains. The study will also include the development and execution of tests to assess the efficiency and effectiveness of the suggested preventative measures. Commonly used in federated learning literature datasets and evaluation metrics will be implemented in the experimental setting.

In conclusion, this study intends to fill a vacuum in the literature by creating effective procedures for preventing data poisoning assaults in federated learning. Contributing to the widespread adoption of privacy-preserving AI technologies, this study improves the safety, honesty, and dependability of machine learning models in federated learning environments.

## 1.2 Research Objectives

The research objectives of this study are as follows:

- The goal of this study is to identify weak spots in federated learning and to comprehend the dynamics of data poisoning assaults. This goal seeks knowledge about the methods of attack, the effects on training, and the features of poisoned data. Understanding the characteristics of data poisoning assaults in the federated learning setting [6] is the goal of this investigation.
- The goal of this research is to create reliable safeguards against data poisoning assaults in federated education. This goal centers on developing cutting-edge methods for detecting and avoiding poisoned data, such as anomaly detection, resilient aggregation, and model verification. The goal is to improve the safety, honesty, and trustworthiness of machine learning models in federated learning environments [5][7].



- For the purpose of measuring how well the planned safeguards work. To achieve this goal, we will devise experiments to test how well the safeguards we've created protect against and respond to data poisoning. To guarantee a thorough evaluation, we will use representative datasets and evaluation measures typically used in federated learning research [8].
- The goal of this study is to fill in the gaps in our understanding of data poisoning assaults on federated learning by adding to the existing literature on this topic. This goal is to address the lack of methods in the literature that are comprehensive enough to deal with data poisoning assaults in the context of federated learning, which is distributed and protects user privacy. This study will help enhance federated learning and its use in privacy-sensitive settings [6][7] by synthesizing and critically evaluating the available knowledge.

If these studies are successful, we'll have a much better idea of how data poisoning attacks in federated learning work and can create safeguards against them. The improved security and dependability of federated learning systems made possible by this work will aid in this study's goal of increasing acceptance and implementation of privacy-preserving machine learning technologies.

# Chapter 2

## Literature Review

In this research, we will discuss medical data security concerns in the federated way of learning. We already know data breach has become a considerable concern for our digital life and medical data is no exception. Every year, thousands of privately owned medical data get stolen by hackers. To prevent this, federated learning has been used for a while. As we know in a federated way, local devices don't share their local data with the global central model. These local clients just share their parameters and thus federated learning gives an effective way of preventing data breaches in a vast network. But this doesn't mean federated networks aren't vulnerable to attacks. It has already been proven many times that federated systems can be attacked with various kinds of methods and makes the classifiers predict false predictions. Many crucial experiments have already been carried out on this issue. We are going to show literature reviews of some of those research below.

Significant improvements in healthcare data analysis in recent years have resulted from the fusion of the Internet of Things (IoT) with deep learning methods. As healthcare data continuously grows at a rapid pace in both amount and complexity, safeguarding data analysis systems has become more vital than ever. For the purpose of this study, we will be looking at this publication [10] to learn more about the methods that have been suggested for long-term use in healthcare data analysis. This research paper [10], describes an IoT-based system for monitoring and analyzing healthcare data using DFL. Deep learning (DL) helps businesses and researchers. Computer vision can spot fraud and produce self-driving cars. FL trains an algorithm on many decentralized edge devices or servers that don't exchange data. New technologies that ensure data privacy, preserve data correctness, and guarantee long-term usage are necessary in light of recent privacy trends and an increase in data breaches across multiple industries. Health data is so sensitive and widespread that it's likely to be hacked or stolen. The research established a framework and method for collecting local training data. IoT devices gather, acquire, and analyze data in this framework. After the global model is presented, healthcare data will be analyzed for problems. Next, IoT device data will be used to develop a local model. When local training is complete, model changes will be sent to a cloud server without personal data. Because the user's data is saved on his device, his privacy is assured. Shared updates contain no relevant information, thus it's useless to target them. Attacks will be weaker. All updates will be averaged for global model training. As it's taught, additional user instances will increase the global model's

performance. Participants' IoT devices will obtain updated software after the global model is trained. Repeating this procedure ensures the model's correctness. FL is the most important component of this system. A cloud server equipped for AI is essential for DFL's worldwide model training. Connected medical devices collect patient information for local model training. In conclusion, the article presents a paradigm for sustainable healthcare data analysis by bringing together deep federated learning and IoT technology. The proposed approach thoroughly addresses challenges such as scalability, energy efficiency, and data privacy by leveraging deep federated learning which can be further researched for more efficient and sustainable healthcare data analysis systems. The experimental results of this paper also provide evidence of the effectiveness of the approach in real-world healthcare datasets.

The field of healthcare has recently shown great potential in Federated Learning (FL) as an effective and fruitful method for collaborative machine learning. This literature review analyzes this study [11] to learn more about the research's findings and architecture plan for FL in the healthcare industry. The authors of this paper [11] conducted a literature review on FL with respect to the application of EHR data in healthcare settings. Solutions, case studies, and ML techniques were provided, and the study uncovered the most important areas of inquiry. The article also covered tips for applying FL to healthcare records. The guidelines of a literature review served as inspiration for this format. Because of this architecture, medical facilities and research institutes can safely use protected health information data in distributed data analysis and machine learning studies. Photographs, lab results, and prescriptions are just a few examples of the kinds of data that "data owners" may keep. It is assumed that data owners understand that joining FL necessitates the use of analysis and learning models that safeguard user privacy. In exchange, the four businesses might potentially hire other private firms to study medical data for insights. Local datasets are never shared directly between their owners due to their confidential nature. Participants in the education process might save electronic health records in a variety of formats and guidelines. In order to train a model, each data owner uses their own unique collection of data. In some cases, this model may make use of data from other forms of FL management. Then, the dataset processing component provides access to and interpretation of local datasets to learning algorithms from outside the system. The module for verifying learning algorithms looks for malicious disclosure of information while using an external algorithm. With the use of local datasets, training models are developed and put into action. Anyone with access to a system that allows for the encrypted transfer of learning models and analysis results via unsecure channels of communication. The manager aggregates all private and public models into one global learning model via Model Aggregation. The manager keeps tabs on the sum of all available public models. It is possible that the model will be given to the student depending on its purpose. In the end, the paper concludes by summarizing the literature on FL in healthcare data and introducing an all-encompassing architecture, both of which add to our knowledge of the topic. Researchers in the healthcare arena can benefit enormously from the findings and proposed architecture, which advocates for FL as a collaborative and privacy-preserving method for analyzing healthcare data.

Since healthcare data is so personal and confidential, its protection is of the utmost

significance. In this literature review, the research paper "Application of Robust Zero-Watermarking Scheme Based on Federated Learning for Securing Healthcare Data." [12] combines robust zero-watermarking with federated learning as a novel approach to healthcare data security and privacy. In [12], the authors discussed how to safeguard healthcare data via federated zero watermarking. Concerns have been expressed concerning IoMT privacy and data security in healthcare. The main idea of this research is to use the Internet of Medical Things (IoMT) in dermatology diagnosis and add a watermark so the patient's information and dataset can't be accessed or shared without permission. In teledermatology, a smartphone picture is utilized to make a remote diagnosis. The dermatological picture might be hacked while being transmitted. This might disseminate patient data. This article showed how federated learning can train a sparse autoencoder network. A trained sparse autoencoder network pulls features from a dermatological picture. Zero-watermarking involves a 2-D Discrete Cosine Transform (2D-DCT). Distributed machine learning system with robust encryption and privacy protection was employed. It aims to allow individuals to train machine learning models without exposing personal data. FL teaches sparse autoencoders how to function together. Second, a trained sparse autoencoder network decodes a smartphone dermatological picture. Last, 2D-DCT is utilized to modify the image's characteristics using low-frequency transform coefficients. Then, a binary feature vector is constructed by comparing the low-frequency coefficients' gray levels to their averages. The watermarking extraction key is created by combining the binary vector and scrambled image. They go in and depart the same way. Dermatology has a bad reputation. Dermatological images can have features extracted from them by the same trained sparse autoencoder network. It can be thought of as adding a watermark while creating a binary feature vector. This robust binary feature vector is cleverly associated with the extraction key used for watermarking. In conclusion, the study makes a substantial contribution to the field of healthcare data security by proposing a robust zero-watermarking system based on federated learning. The framework provides a workable strategy for preserving the confidentiality of individuals' health records. The confidentiality of patients' medical records can be improved with further research and development in this field.

In the healthcare industry, using federated learning (FL) could be a way to get information from disparate sources without compromising privacy. This literature review goes into the study "Federated Learning for Smart Healthcare: A Survey" [13] to examine the in-depth survey of FL's use in futuristic healthcare systems. In [13], the author discussed research issues and future FL research in smart health care. In the past, smart healthcare systems examined health data using AI functions in the cloud or data centers. Due to the rising number of IoMT devices and health data in current healthcare networks, this centralized approach is not successful in terms of communication latency and network scalability. Risks to users' privacy, such as data leakage and breaches, are associated with using a centralized server or third party for data learning. The consumer base is diverse, thus people's blood types, heart rates, facial appearances, and core temperatures will vary. The majority of FL algorithms are tested on a small, generic dataset. Several users of FL-based smart healthcare systems may act maliciously, causing model aggregation to be compromised through the dissemination of harmful model updates or fabricated data. While a model is

being trained locally or shared between clients and the central server, an attacker may tamper with data features. An attacker may assault the server to steal global model data. This is a server-side exploit that leaks private information. FL-based smart healthcare systems have tough security issues. Differential privacy may avert data leaks in training datasets. Secure aggregation methods may be used to encrypt local updates and perform key sharing between clients and the central server. This prevents client assaults and data alterations. The paper's thorough analysis of FL's role in smart healthcare is a significant contribution to the field as a whole. Researchers, practitioners, and policymakers can use the survey's findings to inform their use of FL to safely and securely reap the benefits of dispersed healthcare data. Smart healthcare systems driven by FL will benefit from further research in this area.

Federated Learning (FL), a form of machine learning that puts the privacy of patients first, has lately gained attention in the healthcare sector. This article analyzes the paper [14] which delves deeper into the report's examination of FL's potential applications, barriers, and benefits in the medical field. IoMT devices never disclose personal health information outside of the site or device where it was obtained [14]. In a data-driven society, private data is vital. In medicine, setting up multi-center research is complex since clinicians must decide where to keep data. Minimizing data and limiting its usage are the keys to preventing privacy abuses. Federated learning (FL) is a mechanism for several computers to train an ML model. FL doesn't share data with untrustworthy persons but does share model parameters. Clients may contribute data to be rated and utilize social media to encourage model training. In healthcare, trained ML models should be utilized as trusted guidance. FL security is crucial. This analysis examines how FL for digital health handles privacy and security. The FL articles discuss several strategies, making them hard to appraise. This area lacks FL system norms, which research should alter. This data may be utilized to improve risk models and make healthcare choices. It may also be used to discover trial patients. We believe FL will become the mainstream approach to managing medical data. Privacy and encryption are research concerns. To sum up, this paper [14] is an informative look at the uses, difficulties, and gains of FL in the medical field. The results of this study shed light on how FL can improve healthcare by fostering teamwork in education without compromising patient confidentiality. Healthcare results can be enhanced by improvements in privacy-preserving machine learning, which will be the result of continued research and development in this area.

Research into Federated Learning (FL), a form of collaborative deep learning that enables the training of models by several parties without sharing any raw data, has increased in recent years. However, the safety of FL systems is of paramount importance because of the many threats they face. This article provides a literature review of the study [15] which analyzes the use of GANs to study poisoning assaults in FL.

In the paper [15], a GAN(generative adversarial network) based data poisoning attack is introduced where multiple clients participate in federated learning. Here, the goal was to keep the poisoning task accuracy around 90 percent on average as it will determine how much data the classifier predicted falsely. In this GAN

attack structure, one or multiple attackers can be deployed to create stealth attacks on the global model. The attacker here uses a GAN attack to generate a huge amount of similar samples about different classes of data of different participants. After generating samples from various data classes, the attacker maps them into the locally trained model. The attacker can do the same thing by training his local model on the poison data he has generated and then, when the time is appropriate, uploading the modified parameters to the central server where the global model resides. Because of this, the global model will become tainted, and other locally trained data will be contaminated as a result. In this research, we introduce two neural networks, Discriminator D and Generator G, and describe how GAN operates as an adversarial game between them. In the first, the discriminator is taught to distinguish between the features of the provided data and the features of the created data. The generator is then often taught to sound like the samples utilized by the discriminative networks during training. In this study, we used two different data sets. Both MNIST and ATT(Olivetti dataset) are commonly used. Each image in MNIST is roughly 28x28 pixels in size, while ATT data is 92x112. Both the classifier and discriminator utilized in this image processing project were built using a convolutional neural network-based architecture. The kernel size for the first three convo layers in this CNN model must be 4x4 and for the last convo layer it must be 3x3. Additionally, several advancements have been achieved for these four layers, which are 2, 2, 4, and 1 in order. ReLU and LReLU activation functions have been employed to generate output, and both can be utilized in the discriminator. In one hypothetical attack scenario, performance on the poisoning job first improves to near perfect levels. However, this growth rate will soon decline as more apolitical users submit standard status updates. The average accuracy of the poisoning task can be kept at roughly 90 percent even if this occurs.

To conclude, this paper [15], certainly gives all-encompassing details about how GAN really works and how we can use it for highly effective data poisonous attacks. A malicious attacker can take advantage of the GAN network's Generator and Discriminator functionalities in order to produce poisonous data samples and alter the model.. We can say this paper shows the enormous consequences of adversarial attacks on federated learning as the poisoning task accuracy remains at 90 percent on average during this GAN-based poisonous attack. So, further research in this field is crucial for preserving the security and effectiveness of FL systems.

Adversarial attacks and their defenses have become increasingly vital in the field of deep learning. In order to analyze and create effective defenses against adversarial attacks, researchers need access to varied and difficult datasets. A novel dataset created for assessing the resilience of deep learning models against adversarial attacks is introduced in the research paper "DAMageNet: A Universal Adversarial Dataset," [16] which is the focus of this literature review. An effective way of generating adversarial samples with high transferability has been introduced in this paper [16] where the images can be generated independently and the author named it a zero-query adversarial attack. Here, the original images are called ImageNet, and 96020 of these transferable images or samples are generated by authors that can cheat or mislead many properly-trained DNN networks. Since these adversarial samples have performed really well on black-box attacks and use zero-query procedure, thus

the resulting DNNs perform up to 90 percent error rate which means this attack can make neural network models successfully misclassify 90 images out of 100. The authors called this dataset DAmageNet. For training purposes, varieties of different DNN models have been used on both ImageNet and DAmageNet samples and the results are quite extraordinary. In almost all models, the DAmageNet has over a 90 percent error rate. This shows the true nature of how powerful and deadly adversarial attacks can be for our different deep learning networks. In conclusion, this paper "DAmageNet: A Universal Adversarial Dataset" strengthens the study of adversarial assaults and responses by offering a large-scale, difficult-to-master benchmark dataset. To enhance trustworthiness and security of deep learning models in the real world, as the authors of this study point out, stronger and more robust protection mechanisms are needed. Further study in this field is needed to help improve adversarial defense methods and build more reliable machine learning systems.

Deep Federated Learning (DFL) has recently emerged as a viable method for ensuring data privacy during collaborative machine learning. However, the safety and reliability of FL systems can be jeopardized by malicious attacks on the trained models. This article is a literature review of the study "Analyzing Federated Learning Through an Adversarial Lens," [6] which looks into the dangers that FL systems face from adversaries.

In this paper [6], the authors explored model-poisoning tactics instead of data poisoning. Model poisoning is a poisoning strategy where the poisoning is carried out by an adversary who controls some amount of malicious agents and the number of malicious agents is usually 1. These agents have the ability to make the global model (which is in a central server) misclassify data or inputs with very high confidence. The authors proposed two kinds of model poisoning. One is targeted model poisoning and the other one is stealthy model poisoning. In targeted poisoning, malicious agents typically generate their updates with a simple optimization technique designed to cause a high amount of targeted misclassification. But as there are already large amounts of agents providing updates makes this is even more challenging. Thus, the authors used explicit boosting of the malicious agent's updated parameters. This is done in such a way that it will decrease the combined effects of the other benign agents. On the other hand in stealthy model poisoning, stealth tactics are deployed in malicious weight updates to avoid detection. Although we are aware of the use of explicit boosting in targeted model poisoning, it is possible to detect this boosting across all rounds by employing stealth measurements. So, there is a problem in that our malicious agent can get caught by the global model. For this, stealth model poisoning is carried out where the authors just modify the malicious agents to an extent that they can update their own malicious weights properly without getting detected for most of the rounds. For this experiment, two different datasets were used. The first one is Fashion-MNIST and the second one is the UCI Adult Census dataset. After training the datasets with CNN, the targeted data poisoning attack was quite successful at making the global model give false predictions. After 3 iterations, the global model gave high-confidence predictions and converged with better performance on the validation set. For the stealthy attack, after two iterations, the authors were certain that they had met the adversarial aim for the entire model. This paper also talks about Byzantine resident aggregations [17] that

are not entirely robust to our attack and proved that targeted model poisoning with high confidence was still possible even if the deep neural network uses this aggregation mechanism. Byzantine resilient aggregation’s primary goal is to guarantee convergence to inefficient models, such as those with poor input data classification. Instead of picking one update at random and applying it to the global model, a coordinate-wise median uses all of the relevant data from all of the agents to determine the new update. Therefore, we can say model poisoning attacks are effective against two completely different Byzantine-resilient aggregation mechanisms.

Overall, we can say that this paper quite successfully provides an in-depth analysis of the vulnerabilities and potential adversarial threats faced by federated systems. This paper also emphasizes the urgent need for protection mechanisms such as robust aggregation, and Byzantine fault tolerance. Besides, this paper also describes various model poisoning techniques quite well such as targeted poisoning and stealth poisoning, and shows the downside of explicit boosting which can be bypassed through stealth poisoning methods to avoid detection. Hence, More studies in this area will improve the safety and dependability of FL systems in practical scenarios.

In recent years, the distributed machine learning paradigm known as Federated Learning (FL) has emerged, allowing several users to jointly train models without having to share their raw data. However, the integrity and efficacy of the trained models can be compromised by data poisoning attacks, which target FL systems and include hostile participants injecting poisoned data. Attacks against federated learning systems via data poisoning: an investigation of vulnerabilities and responses is the subject of the research article [18], which is the focus of this literature review.

Many more viable strategies exist for developing poisoning attacks from an antagonistic perspective. The paper [18], demonstrated a powerful strategy for data poisoning in federated learning as well as an effective countermeasure. The authors used what is known as a label-flipping technique to develop an adversarial attack. This attack is particularly effective against FL systems in which a small number of deliberately corrupted participants have been recruited to update the global model with derived mislabeled data in order to poison or corrupt it. Assume a scenario in which an adversarial device controls a portion of the participants in a FL experiment; for simplicity, we’ll state that  $m$  percent of the players are malevolent out of a total of  $P$  participants. The remaining  $m$  percent of good actors will be bribed into poisoning the global model by the  $m$  percent of bad actors. A set number of FL rounds will be used to complete this. The author set out to increase the error rate for select subset classes by manipulating the learnt parameters for the global model designated  $M$ . This method of assault is distinct from those that are not specifically directed. The capacity of these focused attacks to lessen the influence of non-targeted classes of data and hence minimize the likelihood of the poisoned attack being detected is a major benefit of this approach. The authors used two widely-used picture categorization datasets in their study. CIFAR-10 [19] and FashionMNIST [20] are two examples. Each of the 10 categories in CIFAR-10 has 6,000 unique color images, for a total of 60,000. Similarly, Fashion-MNIST is made up of 60,000 photos, with 10,000 images making up each class group. The label-flipping strategy was briefly discussed in this paper. Let’s pretend there’s a  $C(\text{source})$  class



and a  $C(\text{target})$  class. Each malicious actor will change the classification of a subset of the dataset ( $D_i$ ), for example,  $D_i$  data from the  $C(\text{source})$  class will be reclassified as  $C(\text{target})$ . The authors of this work referred to this assault as  $C(\text{source}) \longrightarrow C(\text{target})$ . In scenarios with 2–50 percent malicious individuals, the authors analyzed malevolent accuracy, global model accuracy, and source class recall. When  $m$ , the fraction of hostile agents, hits 40 percent, the recall source classes steadily fall to 0 percent, and the global model accuracy similarly decreases to 74.4 percent from 78.3 percent in dataset CIFAR-10, therefore the attack does fairly well. While the assault fares well on the CIFAR-10 dataset, it is not quite as effective on Fashion-MNIST. Here, the source class recall for the Fashion-MNIST dataset is 58.2 percent with a 30 percent malicious scenario, which is significantly better than the performance of the CIFAR-10 dataset, which is about 19.7 percent. Therefore, the CIFAR-10 dataset is significantly more vulnerable than the Fashion-MNIST.

In conclusion, this paper implemented a different approach when creating an adversarial attack and that is a label-flipping attack. The authors successfully utilized the effectiveness of this attack as the average global model accuracy decreases substantially after attacking both of those used datasets. The results of this study can help researchers and practitioners design better defenses for protecting FL systems from data poisoning attacks.

Recent years have seen an uptick in research on how susceptible machine learning models are to malicious attacks. By manipulating the input data, adversarial assaults aim to deceive machine learning algorithms into classifying data incorrectly. The use of driverless vehicles, image identification, and natural language processing are just a few fields where these attacks could have a disastrous effect. Scientists have suggested a variety of defenses against this hazard. This paper “Efficient Defenses Against Adversarial Attacks” [21], aims to provide an overview of the research on an effective defense against an adversarial attack.

The authors of this paper [21], showed different kinds of adversarial attacks and implemented defenses against them. To begin, several different kinds of assaults have been created, such as Projected Gradient Descent, the Basic Iterative Method, and the Fast Gradient Sign Method. These attack strategies introduce noise into the input data to exploit the model’s flaw. Second, academics have advocated a wide range of protection measures. These consist of gradient regularization, input transformation, and adversarial training. In order to create a more reliable model, adversarial training augments the training dataset with adversarial examples. The purpose of input transformation techniques is to preprocess the input data in order to either completely remove or drastically reduce adversarial perturbations. The decision boundary of the model must be as smooth as possible for gradient regularization techniques to protect against adversarial attacks. Among many other defense tactics, adversarial training paired with input transformation is a successful one. The resilience of the model is enhanced by training it with adversarial examples and simultaneously changing input during inference. The model is strengthened against upcoming attacks by being trained with actual samples of hostile data. Additionally, prior to feeding the input to the model, adjustments like denoising or spatial smoothing can help remove adversarial disruptions. The effectiveness of this in-

tegrated approach in minimizing the harm done by adversarial attacks on various datasets and models has been encouraging. Although adversarial attack defenses have improved, several issues and knowledge gaps still exist. First of all, many of the defense mechanisms currently in use have transferability issues, which means that an adversarial example made for one model can successfully deceive another model. Future research should focus on developing defenses that can be used with a range of models and architectures. The enormous computational cost of current defense strategies must be decreased. For real-time applications, several defenses have excessively large computing costs. We must close this gap and develop strong defenses if we wish to see trustworthy machine learning systems utilized in the real world.

This literature review’s summary of effective defenses against hostile attacks serves as its conclusion. The findings indicate that there are numerous offensive and defensive strategies. Model strengthening techniques that include input transformation and adversarial training have shown potential. However, there are still a lot of issues with transferability and computing overhead that need to be resolved. By filling in these gaps, it will be possible to build defenses that are more effective and practical. In light of the constantly changing adversarial assault scenario, it is more crucial than ever that researchers and practitioners collaborate to develop cutting-edge defenses against machine learning system damage.

Adversarial assaults put the security and dependability of deep learning models in danger. These kinds of attacks involve subtle input sample modifications to produce incorrect categorization or unexpected behavior by taking advantage of model faults. Although various defense strategies have been created in recent years, many of them are either impractically difficult to use or prohibitively expensive. Because it proposes a fresh tactic for repelling hostile attacks, the article [22] is the primary focus of this survey of the relevant literature.

The authors of this paper [22], suggested an effective defense mechanism against adversarial attacks that combines high-level feature representations with denoising methods and increases the robustness of the deep learning models. The authors claim that high-level representations capture the semantic information of the input data, making them more resistant to adversarial perturbations. The proposed method uses these representations as a guide for the denoising process to remove adversarial perturbations and preserve important information. On standard datasets like MNIST and CIFAR-10, the authors evaluate the efficacy of their proposed defense mechanism in comparison to state-of-the-art adversarial attack approaches like the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD). It has been shown experimentally that the proposed method improves the resistance of deep learning models against adversarial attacks, leading to more accuracy and less potential for harm than previous methods. The author’s proposal for a two-stage defense strategy. They begin by utilizing a deep learning model that has been tailored to elicit detailed feature representations. These representations capture important information about the input samples. Second, a denoising module is used to eliminate adversarial perturbations from the high-level representations. The denoising module, which makes use of a trainable denoiser network that learns to differentiate

between the two, eliminates the adversarial perturbations. This research combines a reconstruction loss with a feature loss to create a special loss function for training the denoiser network. While the reconstruction loss evaluates the discrepancy between the denoised and original high-level representations, the feature loss ensures that critical semantic qualities are preserved during the denoising process. The entire denoiser network is trained using adversarial data generated by real-world attack methods. The authors' proposed defense mechanism appears to be making deep learning models more durable to adversarial attacks, while there are still several research gaps that need to be filled. White-box attacks, in which the attacker has full knowledge of the defense mechanism, are the authors' primary concern. When the attacker has no knowledge of the defense, as in a black-box attack, the efficacy of the proposed technique needs more research. The effectiveness of the suggested defense mechanism with larger and more complex datasets should also be looked into.

In conclusion, in order to defend against adversarial attacks, this paper [22] gives a defense mechanism that combines high-level representation extraction with a denoising module. Experimental results show that the proposed approach improves deep learning models' resistance to a wide range of adversarial attacks. However, more research is needed to determine how well the defense mechanism performs in black-box assault scenarios and whether it can scale to larger datasets. Therefore, our study adds to the body of research that attempts to make deep learning models more reliable and safe.

In deep learning, especially when it's used for computer vision, the risk of hostile attacks is rising. Maliciously engineered adversarial perturbations can mislead and undermine the performance of deep learning models, particularly image classifiers. Scientists have looked into a number of different precautions to strengthen the robustness of deep learning models against adversarial attacks. In our analysis of the relevant literature, we focus on this article [23], which investigates the usefulness of using image super-resolution techniques as a defense mechanism against adversarial attacks.

This paper [23] describes how readily deep learning algorithms can be fooled and suggests employing image super-resolution as a countermeasure. The authors conduct a battery of experiments to determine if super-resolution improves adversarial image classification tasks. The results show that the classification accuracy can be improved by using super-resolution techniques to retrieve the original image information and lessen the impact of adversarial perturbations. Here, the authors propose a two-step defense mechanism, with the first stage utilizing a super-resolution network to generate super-resolved images from adversarially altered inputs. The purpose of this procedure is to enhance the resolution of damaged images and restore high-frequency data. Second, the super-resolved images are fed into the image classifier, which improves the predictive power of the model. The proposed method makes efficient use of image super-resolution in order to restore the original image's characteristics and lessens the effect of disruptive interference from enemies. This study fills a huge knowledge vacuum by exploring the potential of image super-resolution as a defense mechanism against adversarial attacks. While the majority

of antecedent research has focused on developing network-level adversarial attack methods and defense strategies, the proposed method offers a novel perspective by employing super-resolution approaches. This knowledge deficit is significant because it emphasizes the need to consider image-level defenses in addition to network-level defenses when attempting to reduce the effects of adversarial attacks. This study explores methods for protecting images from malicious actors, thereby filling a gap in the literature. The proposed method stimulates more study of image-level techniques within the larger subject of adversarial machine learning, and broadens the range of defenses available.

Finally, this analysis of this paper [23], offers useful insights and a potential strategy for protecting against adversarial attacks in classification of Images. The results of this study and the proposed approach pave the way for the creation of more secure deep learning models that can withstand more complex adversarial attacks.

# Chapter 3

## Background Study

In this background study, we examine the dangers associated with federated learning systems. Specifically, we investigate data poisoning attacks and adversarial attacks, both of which can compromise the dependability and effectiveness of trained models. To ensure the safety and efficacy of the proposed system, it is essential to comprehend these attacks and design appropriate defenses. This study lays the groundwork for future research and development of a secured federated learning system for retinal disease identification by investigating the state-of-the-art techniques in federated learning, inception V3, adversarial attacks, and data poisoning attacks, as well as utilizing Retinal OCT datasets.

### 3.1 Federated learning

The use of federated learning has increased in recent years as a response to growing concerns over data privacy and the need for decentralized machine learning. With the advent of federated learning, training machine learning models on dispersed data is facilitated by the proliferation of Internet-connected devices.

One of the primary advantages of federated learning is that it enables businesses to utilize data that would otherwise be restricted due to privacy or regulatory concerns. Federated learning in healthcare allows researchers to train models using data from multiple hospitals to safeguard patient privacy [24]. Similarly, federated learning can be utilized in the financial sector to train models using data from multiple financial institutions without disclosing sensitive data.

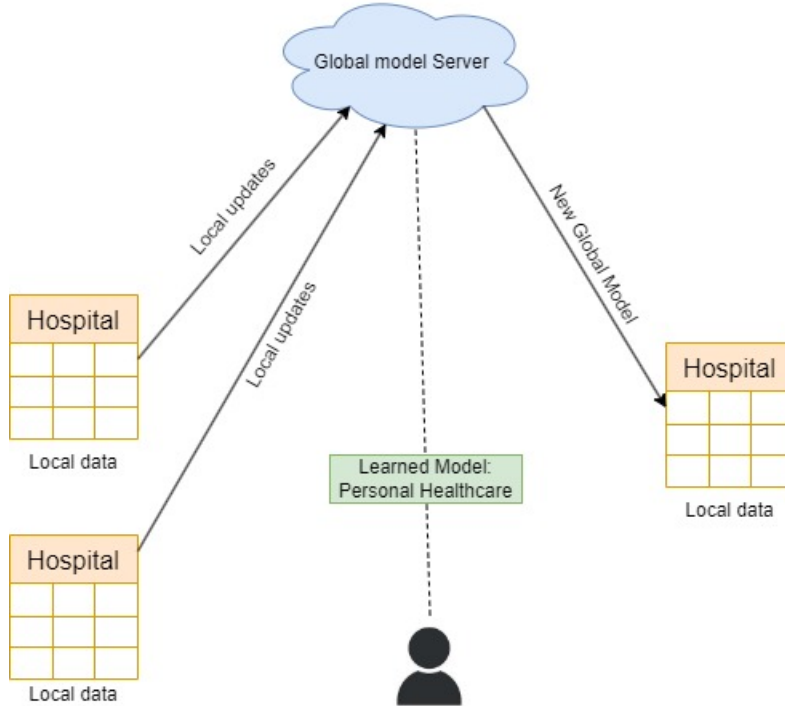


Figure 3.1: Federated Learning Architecture for Hospital Systems

Using federated learning, the time and resources required to train a model can be significantly reduced. Federated learning enables devices to train models locally, eliminating the need to transport massive quantities of data to a centralized server [24], thereby saving time and money. In addition, decentralizing training through the use of locally-hosted models can reduce wasteful time and boost productivity.

Federated learning is a distributed machine learning method in which models are trained locally on endpoints without the transmission of data. The procedure commences with the distribution of an initial model to devices, followed by the execution of local training, the accumulation of updates on a centralised server, and the subsequent iterative refinement of the model. It safeguards user privacy, utilises distributed data, and reinforces models.

$$\mathbf{w}_{\text{avg}} \leftarrow \frac{1}{\sum_{k=1}^K n_k} \sum_{k=1}^K n_k \cdot \mathbf{w}_k$$

However, federated learning is also associated with a number of obstacles. When devices collect data in different methods, for instance, it may be difficult to ensure that the models converge to a consistent global model [24]. Due to the increased device-to-device communication that may be required for federated learning, there is a potential for increased bandwidth requirements and energy consumption.

In conclusion, federated learning is an innovative new direction in machine learning that has the potential to significantly alter the current state of model training and deployment. As research in this area progresses, it is anticipated that new methods and algorithms will arise to improve the efficacy of federated learning.

## 3.2 InceptionV3

In 2015, Google researchers designed the InceptionV3 convolutional neural network (CNN) architecture [25]. This model builds on InceptionV1 and InceptionV2 to enhance image classification accuracy and efficiency. Object and face recognition, as well as image segmentation, are just a few of the numerous computer vision applications that InceptionV3 has facilitated.

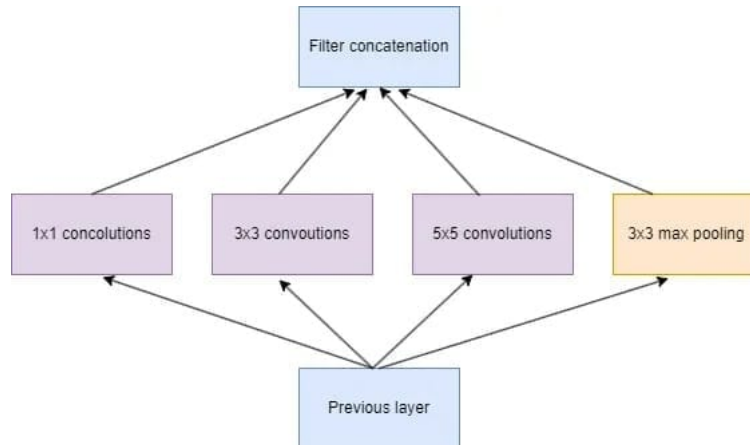


Figure 3.2: Inception V3 Model Architecture Diagram

### 3.2.1 Architecture

The architecture of InceptionV3 is defined by "Inception" modules [25], which are groupings of layers that can simultaneously execute multiple convolution types. In order for the network to learn more complex patterns, these modules integrate features of various sizes and dimensions extracted from the input image. The "Inception-A" block, consisting of four independent routes [25], is the basic element of an Inception module. The first path employs a 1x1 convolution to reduce the number of input channels. In the second method, scale-dependent features are extracted using a 1x1 convolution followed by a 3x3 convolution. The third route utilises a 1x1 convolution followed by a 5x5 convolution to derive features from a larger receptive field. A 3x3 max-pooling operation followed by a 1x1 convolution along the fourth path are used to reduce the output's dimensionality. Each of these branches [25] leads to the subsequent level, where its outputs are combined and utilized as inputs. InceptionV3 contains not only these simpler Inception-A modules, but also the more complex Inception-B and Inception-C modules.

The overall architecture of InceptionV3 is comprised of convolutional layers, Inception modules, a completely linked layer, and a softmax classifier. The output of the softmax layer is a probability distribution over the classes of the dataset.

Along the channel dimension, the Concatenate function concatenates the outputs of these layers.

The InceptionV3 CNN architecture achieves cutting-edge performance in a variety of image classification tasks. Its use of Inception modules enables it to extract features

of varying sizes and dimensions from input images, and its architecture strikes a balance between accuracy and efficiency. As a result, InceptionV3 has become a popular option for a variety of computer vision [25] applications and is likely to continue to be an indispensable tool in this field for the foreseeable future.

### 3.3 Label Poisoning

The objective of adversarial attacks involving label flipping, also known as targeted misclassification attacks [26], is to force a model to make a prediction outside of its intended domain of application. An adversary conducts a label flipping attack by modifying the input data so that the model inaccurately identifies it as belonging to the target class.

There are three stages to a label-flipping attack

- **Crafting adversarial examples:** By making subtle changes to the input data, the attacker generates adversarial samples. These changes, which are too small for human perception, can have a major impact on the model's predictions [26]. Using an optimisation approach, the attacker determines which input perturbations will result in the model incorrectly labeling the input as belonging to the target class.
- **Adding the target label:** Once the hostile examples have been crafted, the attacker modifies the labels of the examples to match the target label. As a result, the model is coerced into incorrectly predicting the target class.
- **Evaluating the attack:** The success of the attack is determined by the attacker's assessment of the model's performance on hostile instances. The assault is successful if the hostile instances have accuracy much lower than the clean examples.

Label flipping assaults can undermine a variety of security features, including anti-spam filters, fraud detectors, and biometric identification systems. Using a label flipping attack, an adversary could deceive a facial recognition system [26] into incorrectly designating him or her as a trusted user by generating a malicious example. Label-flipping attacks are difficult to defend against because they exploit weaknesses in the model's decision-making process. To make the model more resistant to label flipping assaults, however, techniques such as adversarial training and input preprocessing can be utilized.

In conclusion, label-flipping attacks are a form of adversarial attack that attempts to coerce a model into making a prediction for a particular target class. In applications such as biometric authentication systems, these attacks can be used to circumvent security safeguards with catastrophic results. To defend against label-flipping attacks, researchers must develop innovative defense mechanisms and strategies to make the model more resistant to manipulation.

Data poisoning attack is one of the renewed adversarial attacks. The objective of adversarial data poisoning attacks is to modify the inputs to a machine learning



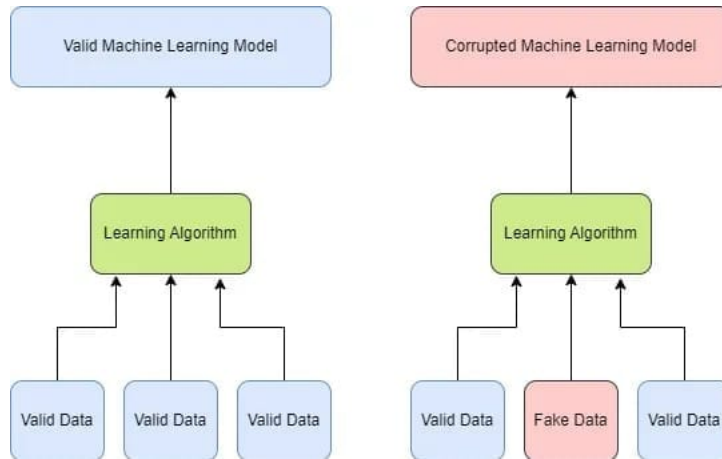


Figure 3.3: Data Poisoning Architecture Diagram

model during its training phase. The objective of data poisoning [27] assaults is to contaminate the training set so that the model generates inaccurate predictions.

The following are the stages of a data poisoning attack:

- Injecting malicious data: The attacker introduces a sample of compromised data into the training set. In order to trick the model into making erroneous predictions, the malicious samples are made to look like authentic data.
- Training the model: Poisonous data [27] is used to train the model. As the model learns from the malicious instances, it will continue to generate inaccurate predictions even when presented with correct data.
- Evaluating the attack: The success of the assault is determined by the accuracy of the model on subsequent data, which is measured by the attacker. The assault is successful if the accuracy of the forecasted data is much lower than that of the clean data.

Spam filters, recommendation systems, and autonomous vehicles are just a few instances in which data poisoning attacks can be used to distort machine learning model results. For instance, a data poisoning attack could be used to trick a recommendation system into advertising harmful products or to cause an autonomous vehicle to make unsafe decisions.

Data poisoning assaults are difficult to defend against because they corrupt the training data for the model. Data cleansing and rigorous training are two methods for making models more resistant to data poisoning attacks. To make a model more resistant to attacks, "robust training" involves training it on a collection of so-called "adversarial examples" and "data sanitization" involves [27] identifying and removing potentially detrimental data from the training set.

Overall, data poisoning assaults are an adversarial attack vector that targets the data used to train machine learning models. Models of machine learning are susceptible to manipulation by these techniques, necessitating the development of novel defense mechanisms and strategies to increase the model's resistance to adversarial attacks.

### 3.4 Data Set

The "Retinal OCT Images" dataset available on Kaggle [28] is a collection of Optical Coherence Tomography (OCT) images of the retina, which is the light-sensitive layer at the back of the eye. The dataset contains a total of 84,495 images, each of size 496 x 768 pixels [28]. They are divided into four categories:

- Normal: OCT scans showing a normal, disease-free retina fall into this category.
- DME (Diabetic Macular Edema): DME, a diabetic condition that leads to fluid accumulation in the macula (the center region of the retina), is depicted in OCT pictures here.
- Drusen: OCT images of a retina with drusen, yellow deposits that can collect under the retina and lead to vision loss, can be found here.
- CNV (Choroidal Neovascularization): CNV (choroidal neovascularization) [28] is a condition in which aberrant blood vessels form beneath the retina, resulting in visual loss, and the corresponding OCT images may be found here.

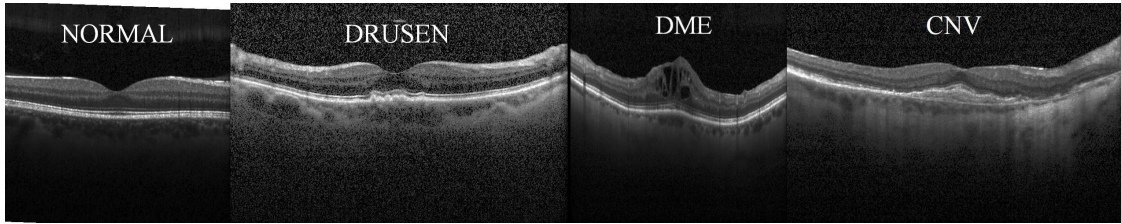


Figure 3.4: Retinal OCT Dataset Samples.

The dataset is useful for creating CAD systems that aid ophthalmologists in the rapid and accurate detection of retinal illnesses using OCT images, and it was collected from retrospective cohorts of adult patients from multiple medical institutions. The images are split into the train, test, and val folders and organized into subfolders for each image category. This organization allows for easy access and use of the data during the training and evaluation processes.

### 3.5 Softmax Function

Activation functions play a crucial role in machine learning and artificial neural networks due to the nonlinearity they impart to the model and the computational complexity they enable. For instance, the softmax activation function is frequently used in applications involving multi-class classification. The outputs of a neural network are transformed into a probability distribution [29] encompassing multiple classes, with weights assigned to each class. The softmax function normalizes the outputs so that their sum equals one, which facilitates their interpretation as probabilities.

**Mathematical Formulation:** The softmax activation function produces a corresponding vector of probabilities when applied to a vector of real-valued inputs [29]. The

softmax function computes the probability that every element in an n-dimensional input vector,  $z$  belongs to a particular class. Here is a description of the softmax function:

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^n e^{z_j}}, \text{ for } i = 1, 2, \dots, n$$

where  $\sigma(z_i)$  is the output of softmax applied to the  $i$ -th item in  $z$ . The exponential of the  $i$ -th element in  $z$  can be calculated using  $e^{z_i}$ , and the sum of all the exponentials of the elements in  $z$  can be calculated using  $\sum_{j=1}^n e^{z_j}$ . The softmax function normalizes the data so that the sum of the output probabilities is 1.

**Benefits and Properties:** The proposed federated learning system deploys the softmax activation method for the detection in the retina due to its many advantageous qualities:

- **Gradient-Friendly:** Softmax is differentiable, gradient-based optimisation techniques like backpropagation [29] can be used effectively. Incorporating a wide variety of optimisation techniques into the training of neural networks is made possible by this trait.
- **Probabilistic Interpretation:** With the probabilities of each class's output that Softmax generates, the model's predictions can be understood as assurance levels. Assigning a probability to each class provides a level of certainty, which is very helpful in multi-class classification problems.
- **Non-Negative Outputs:** The softmax function is useful in situations where decisions need to be made based on probabilities since it guarantees that the output probabilities are positive [29].

Examining the advantages and disadvantages of softmax activation, such as probabilistic interpretation, non-negative outputs, normalization, sensitivity to input differences, and gradient-friendliness, is essential. Softmax may produce less discriminatory results [29] due to its susceptibility to outliers and inability to explicitly reflect class margins. If softmax has difficulty capturing the patterns of minority classes due to a data imbalance, confidence levels may decrease.

This study concludes by demonstrating the potential of a secure federated learning for detecting retinal maladies using OCT images. The paper discusses the disadvantages of centralized deep learning and emphasizes the importance of safeguarding sensitive data. The objective is to increase the robustness and accuracy of the federated learning system through the utilization of confidence scores. The report also acknowledges the dangers associated with federated learning, such as adversarial and data poisoning attacks, and emphasizes the need for robust defenses. The findings of this study establish the groundwork for future work on a secure federated learning system for detecting retinal diseases.

# Chapter 4

## Work Plan

The objective of our study is to use the confidence score as a diagnostic instrument for retinal diseases. Here is the strategy we have in place to achieve this:

### 4.1 Data curation

In this phase, relevant datasets for training and evaluating models for retinal disease are collected. It is essential that the data is sufficiently diverse, representative, and of high quality.

### 4.2 Curation of local data from total data subset

In a federated learning system, data is shared between multiple organisations and/or devices. Here, we will select and curate a subset of the complete dataset for each participant entity. We will use this locally collected data to train models.

### 4.3 Data preprocessing

Data preparation is the initial step in preparing data for training purposes. Data cleansing, missing value management, feature engineering, and normalisation are all conceivable steps. This process prepares the data for model training and ensures that it is in the proper format.

### 4.4 Apply data poisoning

Data poisoning is a technique employed by cyber-criminals to introduce false or skewed data into a training dataset. Data poisoning aims to simulate potential assault scenarios against a federated learning system. This checkpoint enables you to evaluate the safety and stability of your proposed system.

## **4.5 Simulate a regular federated learning system and a secure federated learning system**

We utilize the locally selected data by training models on them. Then, we will compare the outcomes of two training scenarios: one in which confidence ratings are used to diagnose retinal disorders without the proposed method (the baseline), and another in which confidence ratings are used to train. The proposed solution is believed to enhance the dependability and security of the federated learning system.

## **4.6 Model evaluation**

After training the models, we will evaluate their performance using the appropriate metrics. This was accomplished by employing the softmax activation method. At this juncture, we can evaluate the effectiveness of the proposed technique in detecting retinal disorders.

## **4.7 Compare results**

Finally, we will compare the results of the training situation (where the proposed method was not implemented) and the scenario where it was implemented. By comparing the two systems, we can determine the impact that confidence scores have on the federated learning system. We will dissect and investigate how the suggested technique improved precision, safety, and overall performance.

At the end, we intend to develop a secure federated learning system for diagnosing retinal disorders. By utilizing confidence scores and testing the system's resistance to data poisoning attempts, we expect to improve the system's accuracy and safety, thereby advancing research into the identification of retinal diseases.

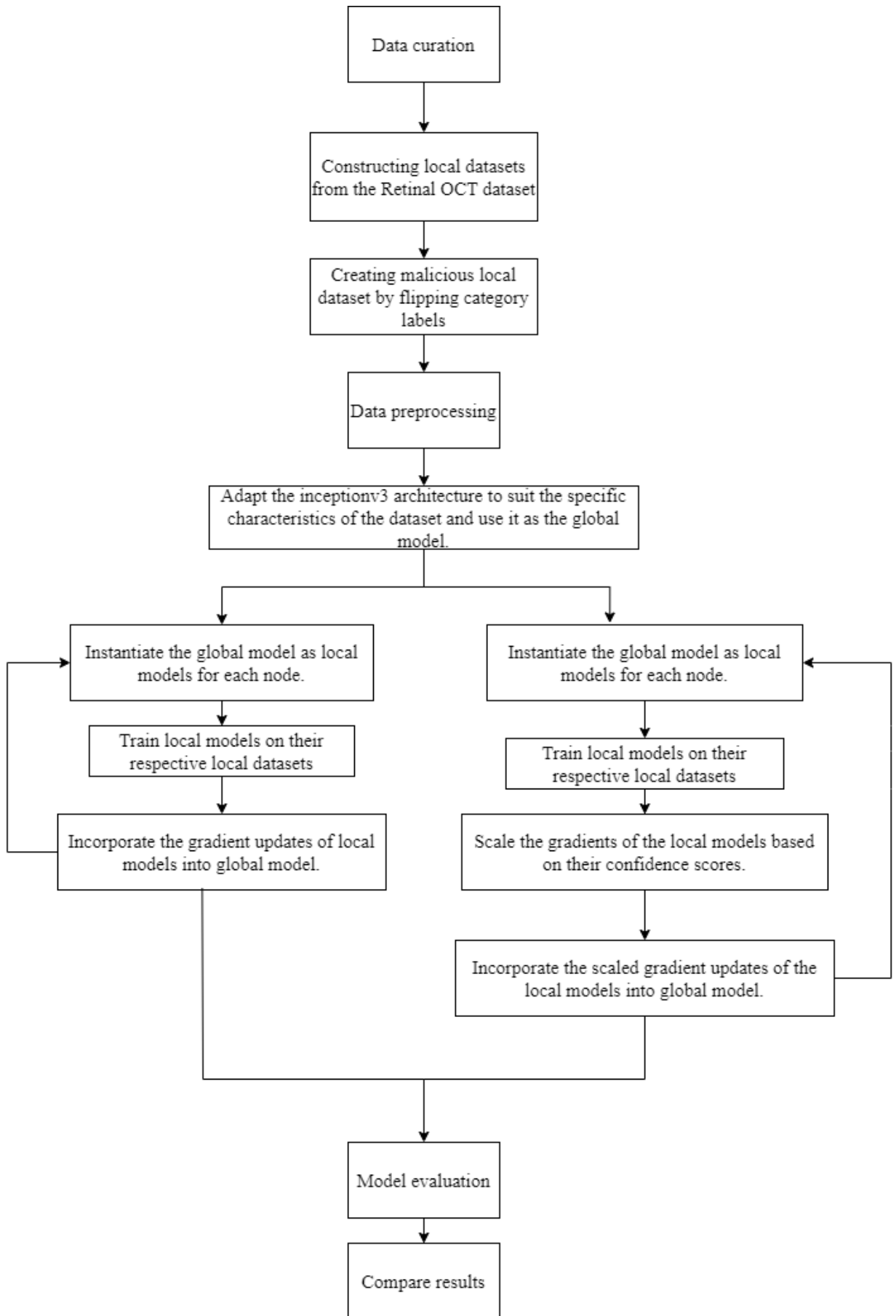


Figure 4.1: Working plan

# Chapter 5

## Proposed Method and Implementation

This section first presents the proposed method in detail, then describes the implementation of our work for its evaluation.

### 5.1 Confidence Score Based Aggregation

The proposed method aggregates local models' gradient updates into the global model by scaling them according to their confidence, which enhances the sophistication of the approach.

To measure the model's confidence score  $C_i$ , we use the average of the SoftMax probabilities [sigma] for the expected category index  $i$  of each input vector  $z$ .

$$C_i = \bar{\sigma}(z)_i$$

After calculating the confidence score  $C_i$  of each local model and normalizing it, we use the normalized score to scale its respective local model's gradient  $w_i$ . Finally, the global model receives the scaled gradient updates from all the local models.

$$w_g = \sum_{i=0}^N \left( \frac{C_i}{\sum C_i} * w_i \right)$$

### 5.2 Implementation

#### 5.2.1 Data Organization

For the experiment, we used a subset of the "Retinal OCT" dataset that contained 9 local datasets. Each local dataset had 2000 images with balanced categories. We intentionally mislabelled a portion of the images in 3 of the local datasets.

	Correct data	Mislabelled data
Malicious dataset 1	1000	1000
Malicious dataset 2	800	1200
Malicious dataset 3	0	2000

Table 5.1: Malicious local datasets



Figure 5.1: Visualization of Local datasets

Table 5.1 shows how much of the local datasets are mislabelled. It can be observed that a substantial proportion of mislabelled data has been used, which provides us with more clear and presentable evaluation results.

Each local dataset was then split into train and validation sets, validation sets containing 10% of the local dataset. Two datasets were created as test sets. One test set containing 400 images is used to measure the confidence score of the local models. The other test set containing 968 images is used to evaluate the global model’s performance.

## 5.2.2 Data Pre-processing

The images are resized into 150x150 pixels and normalized before using them on the model.

## 5.2.3 InceptionV3

InceptionV3 is used as the base model of the federated learning system. We took a topless inception model and added a fully connected layer consisting of 4 neurons,



one for each category. We applied the SoftMax activation function to the output layer. We only trained the top layer, keeping the rest of the model fixed. We chose Adam as the optimizer for our model.

#### **5.2.4 Secure Federated Learning System**

We initialized the global model with InceptionV3 and created nine instances of it as local models. Each local model was trained and validated on its own dataset for seven epochs. After that, the confidence score of each model is evaluated. Then, the gradients of the models are aggregated according to their confidence scores before the global model receives them as updates. This concludes one cycle of the Federated Learning System. We repeated this process for 25 cycles.

For comparison, we also built another federated learning system that did not use confidence scores. It simply averaged the gradients of the local models and updated the global model directly.

#### **5.2.5 Evaluation**

After running the FL cycles, we evaluated both global models to demonstrate an in-depth comparison between them. We used several evaluation metrics such as Accuracy, loss, precision, recall, F1 score, and AUC-ROC to highlight the effectiveness of the proposed method.

# Chapter 6

## Result Analysis

In this section of the paper, we illustrate various evaluation charts and scores.

### 6.1 Confidence Scores

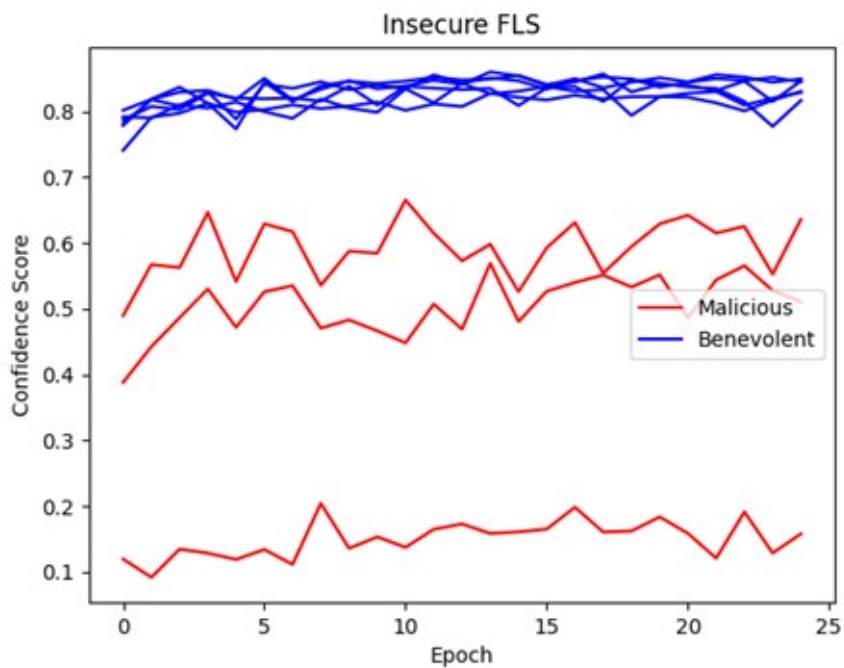


Figure 6.1: Confidence score rates of the Insecure FLS.

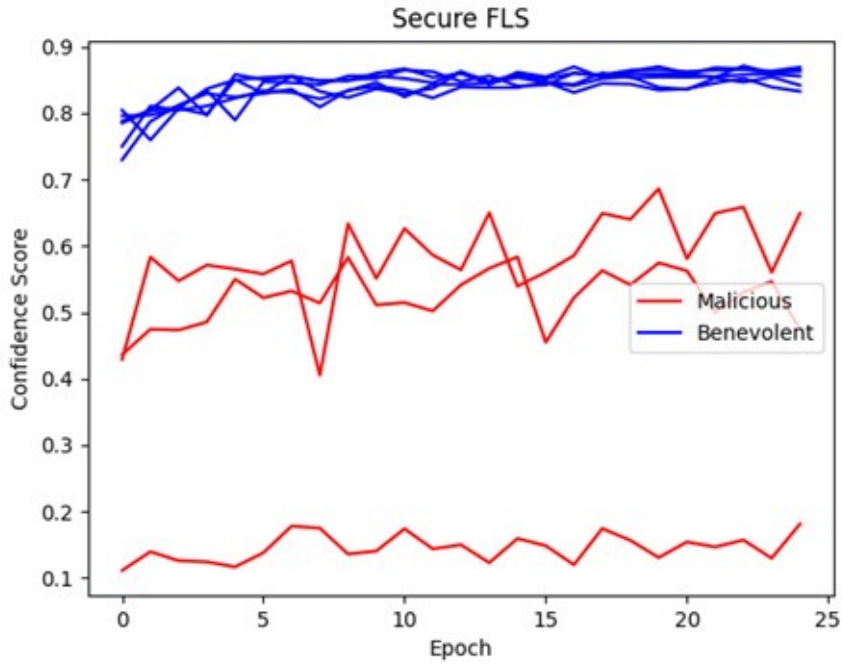


Figure 6.2: Confidence score rates of the Secure FLS.

Figures 6.1 and 6.2 compares the confidence scores of the local models for secure and insecure federated learning systems (FLS) over 25 cycles. It shows two graphs, one for each type of FLS. The local models are either benevolent or malicious, depending on whether they cooperate or interfere with the global model. The difference in confidence scores between the malicious and benevolent models are quite apparent. The confidence score rates of the benevolent local models are closely clustered together. Moreover, the benevolent local models have more stable confidence scores in the secure FLS than in the insecure FLS. In contrast, the malicious local models have lower and more fluctuating confidence scores in the secure FLS than in the insecure FLS. This is because the malicious local models try to fit the global model with different types of datasets, which are more effective in the insecure FLS and less effective in the secure FLS.

## 6.2 Accuracy and Loss

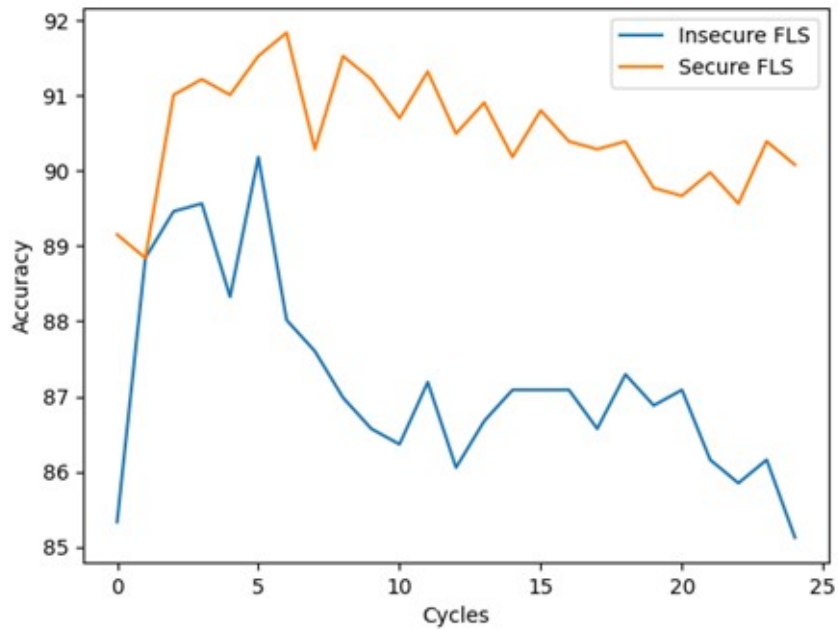


Figure 6.3: Accuracy rates of the Insecure FLS.

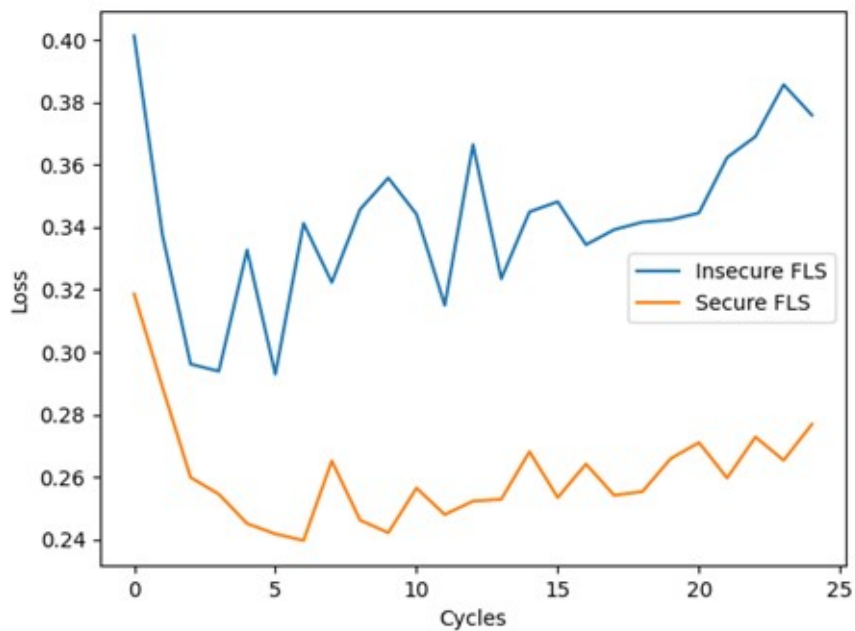


Figure 6.4: Loss rates of the Secure FLS.

Figure 6.3 shows a line chart comparing the accuracy of the global model for secure and insecure federated learning systems (FLS) over 25 cycles. The secure FLS starts

with an accuracy of about 89%, while the insecure FLS starts with about 85%. The insecure FLS reaches a peak accuracy of slightly above 90% before declining, while the secure FLS maintains an accuracy of around 92%. The chart clearly demonstrates that the secure FLS has a higher and more stable accuracy than the insecure FLS. Similarly, Figure 6.4 illustrates that the secure FLS has a lower and more stable loss rate than the insecure FLS.

### 6.3 Confusion Matrix

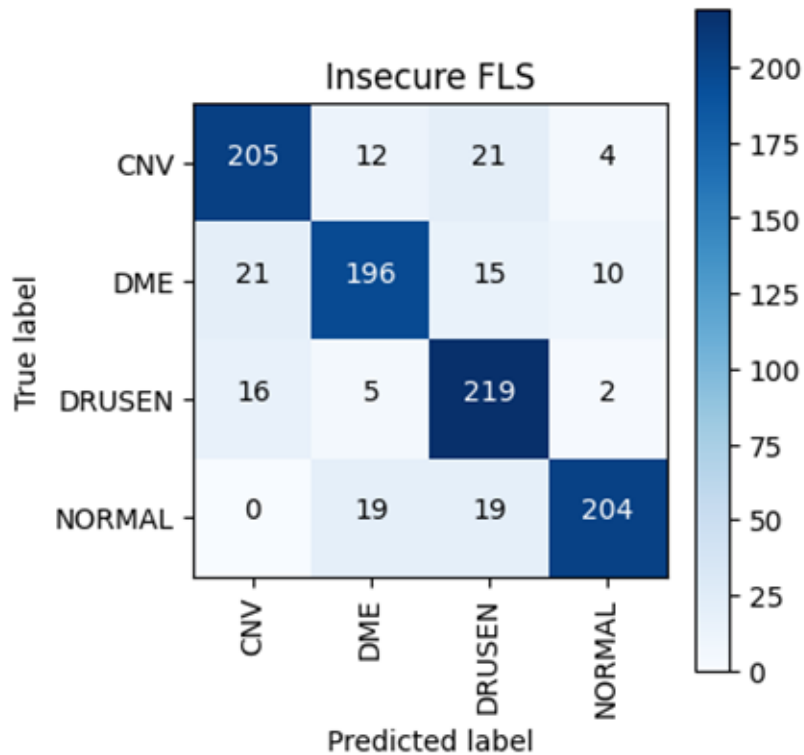


Figure 6.5: Confusion matrix of the Insecure FLS.

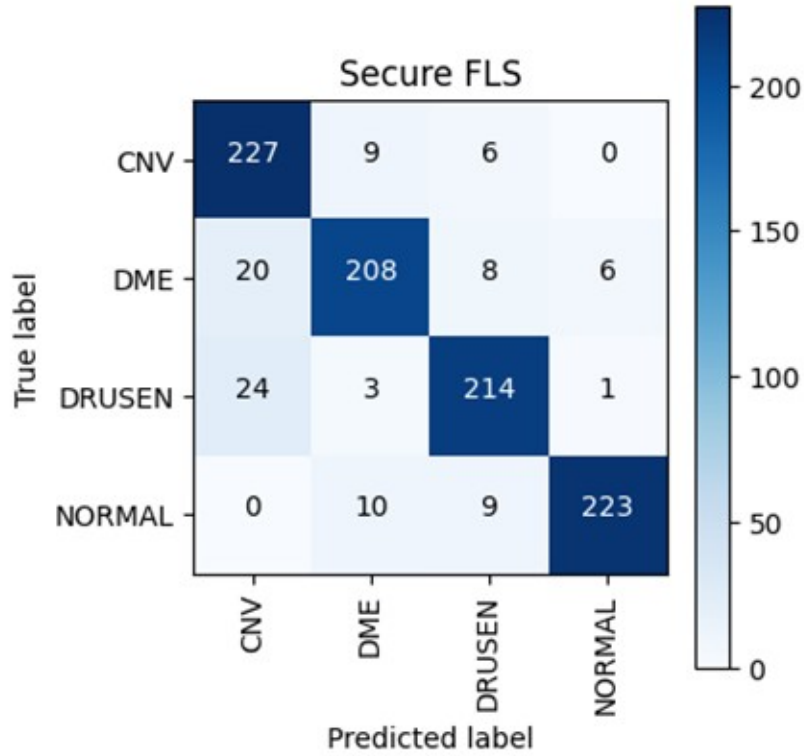


Figure 6.6: Confusion matrix of the Secure FLS.

The confusion matrix of the two systems in figure 6.5 and figure 6.6 provides less clarity to the comparison. However, the secure FLS seems to have better overall results than the insecure FLS. CNV, DME and NORMAL categories has more true positives in the secure FLS, while only DRUSEN category has slightly more true positives in the insecure FLS. To evaluate the confusion matrices better, we can use the following metrics.

Table 6.1 contains the precision, recall, F1 score, and AUC-ROC of insecure FLS and secure FLS. Secure FLS achieves noticeably better score in every metrics.

	Insecure FLS	Secure FLS
Precision	0.8689581	0.9063158
Recall	0.8357438	0.8894628
F1	0.852027359854591	0.8978102417518544
AUC-ROC	0.8357438	0.8894628

Table 6.1: Comparison of performance metrics between Insecure FLS and secure FLS

# Chapter 7

## Conclusion and Future Works

### 7.1 Conclusion

In conclusion, this thesis paper concludes by resolving an urgent issue: how to construct a reliable federated learning system that uses confidence ratings to diagnose retinal disorders. The primary objective of the study was to discover a way to improve the speed and accuracy of diagnosing retinal diseases without compromising data privacy or security. We analyzed the disadvantages of conventional, top-down methods of education and recognised the potential of federated learning in the health sciences. In this study, the Retinal optical coherence tomography (OCT) dataset was subjected to data poisoning, a machine learning adversarial attack technique. During a data poisoning attack, the accuracy of the primary server is lowered. To diagnose retinal disorders, we will therefore employ a technique based on a protected federated learning system that utilises a confidence score. The use of this will not compromise accuracy and will help mitigate damage caused by malicious users. SoftMax values are employed to assess the dependability of local updates and serve as the foundation for the confidence score. Using a Retinal (OCT) dataset and a data poisoning attack, we evaluate our approach. In our federated learning network, there are nine local models, and three of them are poor. The proposed method significantly improves the precision, recall, F1 score, and AUC-ROC of the global model. Specifically, it elevates the precision from 0.869 to 0.906, the recall from 0.836 to 0.889, the F1 score from 0.852 to 0.898, and the AUC-ROC from 0.889 to 0.889 percent, respectively.

### 7.2 Future Works

Our model performed exceptionally well with the provided data. Nonetheless, we are eager to observe how well our model performs when provided additional data, such as a massive number of datasets. If a client possesses only a small amount of potentially detrimental information, it will not cause significant harm. In the case of numerous clients, however, we must investigate the effects of a large volume of detrimental data. If we could conduct a simulated attack against a realistic target, such as a hospital, it would be simpler to determine its effectiveness. Furthermore, we can utilise AI that can be explained. This study paves the way for future developments in federated learning, which will increase trust and collaboration among healthcare professionals and ultimately result in improved patient care and outcomes.

# Bibliography

- [1] Nam, C. W. (2020, July). World economic outlook for 2020 and 2021. In CESifo Forum (Vol. 21, No. 2, pp. 58-59). Institut für Wirtschaftsforschung (Ifo).
- [2] Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., McDaniel, P. (2017). Ensemble adversarial training: Attacks and defenses. arXiv preprint arXiv:1705.07204.
- [3] Li, Y., et al. (2019). False negative adversarial attacks on deep learning-based medical image classification. *Journal of Biomedical Informatics*, 94, 103173.
- [4] Anyaduba, C. A. (2021). *Childhood in contemporary diasporic African literature: memories and futures past*: by Christopher EW Ouma, Cham, Switzerland: Palgrave Macmillan, 2020, 202 pp., ISBN 978-3-030-36256-0 (eBook).
- [5] Tramèr, F., Papernot, N., Goodfellow, I., Boneh, D., McDaniel, P. (2017). The space of transferable adversarial examples. arXiv preprint arXiv:1704.03453.
- [6] Bhagoji, A. N., Chakraborty, S., Mittal, P., Calo, S. (2018). Analyzing federated learning through an adversarial lens. CoRR. arXiv preprint arXiv:1811.12470.
- [7] Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., Shmatikov, V. (2020, June). How to backdoor federated learning. In *International Conference on Artificial Intelligence and Statistics* (pp. 2938-2948). PMLR.
- [8] Yuan, X., He, P., Zhu, Q., Li, X. (2019). Adversarial examples: Attacks and defenses for deep learning. *IEEE transactions on neural networks and learning systems*, 30(9), 2805-2824.
- [9] Kurshan, E., Shen, H. (2020). Graph computing for financial crime and fraud detection: Trends, challenges and outlook. *International Journal of Semantic Computing*, 14(04), 565-589.
- [10] Elayan, H., Aloqaily, M., Guizani, M. (2021). Sustainability of healthcare data analysis iot-based systems using deep federated learning. *IEEE Internet of Things Journal*, 9(10), 7338-7346.
- [11] Antunes, R. S., André da Costa, C., Küderle, A., Yari, I. A., Eskofier, B. (2022). Federated learning for healthcare: Systematic review and architecture proposal. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(4), 1-23.



- [12] Han, B., Jhaveri, R., Wang, H., Qiao, D., Du, J. (2021). Application of robust zero-watermarking scheme based on federated learning for securing the healthcare data. *IEEE journal of biomedical and health informatics*.
- [13] D. C. Nguyen, Q.-V. Pham, P. N. Pathirana, et al., “Federated learning for smart healthcare: A survey,” *ACM Computing Surveys (CSUR)*, vol. 55, no. 3, pp. 1–37, 2022.
- [14] B. Pfitzner, N. Steckhan, and B. Arnrich, “Federated learning in a medical context: A systematic literature review,” *ACM Transactions on Internet Technology (TOIT)*, vol. 21, no. 2, pp. 1–31, 2021.
- [15] Zhang, J., Chen, J., Wu, D., Chen, B., Yu, S. (2019, August). Poisoning attack in federated learning using generative adversarial nets. In *2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And 24 Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)* (pp. 374-380). IEEE.
- [16] Chen, S., Huang, X., He, Z., Sun, C. (2019). DAmageNet: a universal adversarial dataset. *arXiv preprint arXiv:1912.07160*.
- [17] Blanchard, P., El Mhamdi, E. M., Guerraoui, R., and Stainer, J. Machine learning with adversaries: Byzantine tolerant gradient descent. *Advances in Neural Information Processing Systems*, 2017.
- [18] Tolpegin, V., Truex, S., Gursoy, M. E., Liu, L. (2020, September). Data poisoning attacks against federated learning systems. In *European Symposium on Research in Computer Security* (pp. 480-501). Springer, Cham.
- [19] Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009).
- [20] Xiao, H., Rasul, K., Vollgraf, R.: Fashion-mist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747* (2017).
- [21] Zantedeschi, V., Nicolae, M. I., Rawat, A. (2017, November). Efficient defenses against adversarial attacks. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security* (pp. 39-49).
- [22] Liao, F., Liang, M., Dong, Y., Pang, T., Hu, X., Zhu, J. (2018). Defense against adversarial attacks using high-level representation guided denoiser. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1778-1787).
- [23] Mustafa, A., Khan, S. H., Hayat, M., Shen, J., Shao, L. (2019). Image super-resolution as a defense against adversarial attacks. *IEEE Transactions on Image Processing*, 29, 1711-1724.
- [24] McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B. A. (2017, April). Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics* (pp. 1273-1282). PMLR.

- [25] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2818-2826).
- [26] Kurakin, A., Goodfellow, I., Bengio, S. (2016). Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236.
- [27] Gu, T., Dolan-Gavitt, B., Garg, S. (2017). Badnets: Identifying vulnerabilities in the machine learning model supply chain. arXiv preprint arXiv:1708.06733.
- [28] “Retinal OCT Images (Optical Coherence Tomography).” Retinal OCT Images (Optical Coherence Tomography) | Kaggle, /datasets/paultimothymooney/kermany2018. Accessed 13 Jan. 2023.
- [29] Hinton, G. E., Osindero, S., Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, 18(7), 1527-1554.