Secure Computation and Proportionally Fair Collaboration in Federated Learning of Histopathology Images

by

Seyedeh Maryam Hosseini

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Systems Design Engineering

Waterloo, Ontario, Canada, 2023

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Member:	Metin N. Gurcan Director, Center for Biomedical Informatics, Wake Forest School of Medicine
Supervisor:	Hamid Reza Tizhoosh Professor, Rhazes Lab, Artificial Intelligence and Informatics, Mayo Clinic, Rochester, MN, USA
Internal Member:	Otman Basir Professor, Dept. of Electrical and Computer Engineering, University of Waterloo
	Nima Maftoon Assistant professor, Dept. of Systems Design Engineering, University of Waterloo
Internal-External Member:	Denise Hileeto Clinical Associate Professor, Dept. of Health Science, University of Waterloo
Co-Supervisor:	Morteza Babaie Adjunct Assistant Professor, Systems Design Engineering, University of Waterloo

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

The prediction power of deep learning models depends on the size and quality of the training data. Having access to large-scale datasets enables the model to more precisely estimate the underlying distribution of the data. Deep models rely on training datasets that are ideally aggregated from various sources. However, it may not be possible to construct large-scale datasets in one central location in the medical domain due to privacy considerations. In centralized learning methods for medical imaging, training data is supposed to originate from different medical centers (i.e., hospitals and clinics), and be transferred to a centralized location, commonly called a *server*. However, hospitals are generally not willing to share their medical records with other external collaborators because of privacy considerations and regulatory compliance. Therefore, the lack of publicly available large-scale diverse datasets hinders model development in healthcare. To overcome these challenges, decentralized learning methods are a promising scheme to preserve data privacy while enabling training of general models using data from different sources. Federated learning allows training on multi-site datasets without requiring direct access to data. Federated learning has emerged as a promising solution to protect user-sensitive data by keeping data local. It is a novel decentralized paradigm that plays a critical role in privacy-sensitive applications, opening new horizons for secured decentralized learning methods.

The main focus of this research is privacy-preserving federated learning. The two key challenges in federated learning, namely privacy of the training results and fairness in aggregating training results will be addressed. The first challenge is that training results are as important as training samples as they may reveal privacy clues. To address this challenge, this thesis adopts *secure multi-party computation* and proposes a framework enabling participant hospitals to maintain privacy while sharing their training results. The second challenge is that the collaboratively learned global model is supposed to have acceptable performance for the individual sites. However, existing methods focus on model averaging, leading to a biased model that performs perfectly for some hospitals while exhibiting undesirable performance for other sites due to non-iid data distribution among hospitals. This challenge will be addressed by improving the model fairness among participating hospitals through introduction of a novel federated learning scheme called *Proportionally Fair Federated Learning*, Prop-FFL. Proportional fairness modifies the aggregation rule at the central server to account for varying site contributions. It is based on a novel optimization objective function to decrease the performance variation among hospitals. Experiments have been conducted on The Cancer Genome Atlas (TCGA), a publicly available repository. The experimental results suggest competitive performance compared to the baseline and benchmark schemes.

Acknowledgements

Writing this thesis has been fascinating and rewarding. I would like to thank all the people who made this thesis possible.

First and foremost, I am extremely grateful to my supervisor, Professor Hamid R. Tizhoosh for his constant support, encouragement, and patience during my PhD study. I am glad to be part of Kimia Lab, where I had freedom to explore creative ideas.

I would also like to thank Professor Gurcan, Professor Basir, Professor Maftoon, and Professor Hileeto for taking out their time to review my thesis and provide valuable suggestions.

Last but not least, I would like to thank my parents for raising me to value the education. I deeply appreciate my sister and my brother for their infinite support and guidance.

Dedication

To my beloved parents.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Histopathology is regarded as the gold standard for diagnosis of many diseases, among other cancer, and determining the presence and nature of a disease. However, diagnosis by pathologists is subjective with high *intra-and inter-observer variability*. Observer variability is defined as the variation between the diagnosis of a set of cases examined by independent pathologists [1]. Diagnostic observer variability is inevitable due to reasons such as pathologists' educational bias, the complexity of the task, and the continuous evolution of diagnostic criteria [2]. These factors can lead to inconsistent diagnostic interpretation and hinder reaching diagnostic consensus. Over the past few years, the rapid development of artificial intelligence and machine learning has revolutionized the research in healthcare domain. In this regard, digital pathology has offered an excellent opportunity to improve consensus and increase the precision of diagnostics through the adaptation of machine learning techniques, especially deep learning [3]. The key enabler of digital pathology is data. Training a generalized machine-learning model requires collecting data from various medical institutions. However, compared to other domains, medical data is highly sensitive, and collecting data from medical centers in one centralized location seems to be wishful thinking at the present time. Sharing medical data with other third parties raises privacy concerns as the training data may contain private patient information. Additionally, centralized data may not only expose individuals to privacy risks but also faces organizations to legal risks. There are various guidelines and regulations that prohibit medical centers from patients' data disclosure, such as United States Health Insurance Portability and Accountability Act (HIPPA) [4].

Thus, it is absolutely indispensable to develop privacy-enhanced computerized approaches in the healthcare domain to protect both patients and medical organizations. In this context, a crucial question that will be answered in this thesis is *how to train a*

high-performance machine learning model on multi-institutional data without violating the privacy of the patient's data.

Recently the concept of "*federated learning*" has been introduced to enhance the privacy aspect of machine learning methods. The initial work of federated learning has been proposed for mobile devices. Soon after, federated learning was expanded for use in a variety of applications. It has proven its potential not only in wireless communication, but also in the Internet of Things (IoT), autonomous driving, and healthcare, to name a few. Currently, federated learning is broadened to be included across organizations such as hospitals. Federated learning is a distributed machine learning framework that allows for collaborative model training by coordinating tasks among hospitals without sharing actual training samples. More specifically, the central server initializes the global deep model with random parameters first and subsequently sending those parameters to each hospital. Then, each hospital trains the model with its local data, updating the model parameters, and sending local updates – and not the data – to the central server. Next, the central server combines all local model updates and constructs a new improved global model, sharing the new model with hospitals to locally train again. This process continues until the global model converges to a stable solution according to common learning rules.

Federated learning can also be applied without any central server, which is called "*decentralized federated learning*". Both centralized and decentralized federated learning form a collaborative learning environment among medical centers to facilitate privacy-protection during model training by not revealing patients' private data. Federated learning with its innovative operational procedure offers not only data privacy enhancement, but also helps to reduce the need for central storage. It enables us to benefit from diverse datasets from various medical centers representing a much larger diversity of patient demographics. Therefore, there is a higher chance that the trained model be of higher generalization and expectedly achieve higher accuracy, which might not be achieved by conventional approaches with insufficient data.

This Ph.D. research is premised on the hypothesis that federated learning can potentially remedy intra-and inter-observer variability and improve consensus building. In other words, the model learned through federated learning can be used as a "second opinion". This can help pathologists to benefit from the collective wisdom of diagnostic patterns of previous cases interpreted by other pathologists. However, applying federated learning techniques to real-world data may suffer from some challenges, especially of the non-IID nature of the real-world datasets. This thesis will address two challenges of federated learning in the medical domain, namely secure computation and fairness.

1.1 Motivation

The motivation for this Ph.D. research originates from data privacy and data sharing concerns in machine learning. These concerns are more crucial in the medical domain where medical centers and healthcare providers have restrictions around sharing patients' data with external collaborators. At the same time, training a machine learning model for healthcare purposes often requires comprehensive data from multiple medical institutions. Therefore, the question that remains to be addressed is *how to train machine learning models when we might not be able to access training data*. In this regard, decentralized machine learning methods such as federated learning belong to emerging technologies.

Federated learning, as a privacy-preserving framework, involves training a model without having direct access to patient-sensitive data but rather in a distributed collaborative fashion. The federated approach relies on decentralized data distribution from various hospitals and clinics. However, employing federated learning in the medical domain, e.g., for histopathology images, may be impeded by some challenges. One challenge is how effectively aggregate training results to obtain a high-performance model when data is non-IID. The second challenge is how to create a secure communication to protect training results from disclosure. This thesis will propose solutions to address these challenges in federated learning.

1.2 Thesis Objectives and Contributions

The main objective of this thesis is to develop practical decentralized learning frameworks to maintain the privacy of whole-slide images (WSIs) during the training of histopathology data, and to create a secure communication procedure. These frameworks allow decentralized model training, enabling medical centers to collaboratively learn a *fair* model based on security and privacy considerations. The experiments within this research were designed to quantify the performance of the proposed decentralized machine learning methods. To some extent, this thesis contributes to the key challenge of data privacy in the biomedical community, enabling machine learning techniques to be safely applied for histopathology diagnosis of diseases, particularly cancer.

The two significant contributions of the thesis are as follows:

- In Chapter 3, this thesis introduces cluster-based secure multi-party computation (SMC) framework to preserve privacy when participant medical centers and the central server communicate.

- In Chapter 4, this thesis introduces Prop-FFL, a federated learning framework that enables hospitals to collaboratively learn a fair model across medical centers.

1.3 Thesis Organization

The thesis is organized into five Chapters which are structured as follows:

- Chapter 2 introduces the general idea of computational pathology and federated learning. It presents the necessary definitions, concepts, and mathematical formulations of federated learning. Also, the state-of-the-art benchmarks in federated learning will be reviewed.
- Chapter 3 focuses on privacy-preserving methods in federated learning. It presents the proposed secure multi-party computation (SMC) framework to protect privacy in federated learning.
- Chapter 4 presents fairness in federated learning. It introduces the proposed proportionally fair hospital collaboration in federated learning.
- The summary of the thesis along with a general conclusions and future works are provided in Chapter 5

Chapter 2

Background

2.1 Digital Pathology

Medical practice has been revolutionized over the past few decades thanks to emerging of new hardware and computerized technologies. The advent of digitized whole-slide images (WSIs) is about to transform traditional histopathology into a modern field described as "*digital pathology*". A WSI is the digital output of scanning glass slides (with mounted/fixed tissue sample) with high resolution and color depth. It is a digital representation of a glass slide and is sometimes called a virtual microscopy image [5]. WSIs can be saved in computer storage and viewed/processed with appropriate tools and software. Pathologists' main task involves diagnosing cases by accurate interpretation of inspecting glass slides under a microscopic. With the advent of WSIs, pathologists can assess digitized glass slides on a computer monitor instead of examining glass slides through a microscope, i.e., digitized WSIs can fully replace light microscopy in many fileds. Today's modern clinical practice has started to rely on digital pathology for technological requirements in laboratory environments [6, 7]. Digital pathology has occupied such an important place in research of histopathology speciemens that is next to impossible to imagine a future without it. WSI viewing software enables pathologists to annotate, edit, highlight, analyze, and easily share their findings with other experts [8]. As a result, it facilitates clinical workflow and speeds up the diagnostic process. Having everything on a computer helps pathologist to access previously diagnosed cases quickly and easily. Using digital pathology, the computational pathology can speed up the process as it automatically provides us with quantitative analysis, such as nuclei cell and mitosis count. The benefits of using digital pathology are broad and not limited to improving the pathologists' workflow. For example,

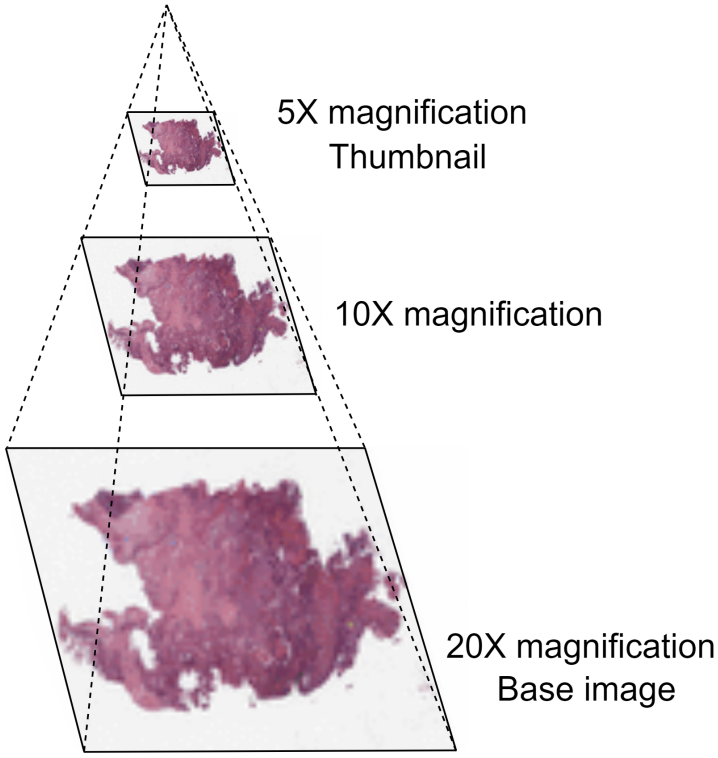


Figure 2.1: An example of WSI format in a pyramid structure. Due to their gigapixel size, WSIs are saved in a multi-magnification structure.

it might resolve storage concerns in laboratories. Currently, medical centers must have a physical space to keep millions of glass slides for several years before they can destroy slides because of space limit [9]. Using WSIs not only relieves the need for physical storage space but also keeps slides available for a longer period. Another benefit is that digital pathology supports pathologists in difficult cases as it allows for remote consultation with internal and external experts and specialists across the world. This would improve the quality of diagnosis and lead to more accurate decisions. Additionally, digital pathology facilitates the research. Having WSIs enables us to apply artificial intelligence and machine learning analysis to histopathology images, a new horizon in computational pathology.

2.1.1 WSI Format

WSIs are extremely large images compared to other medical image modalities such as radiology images. They are often captured in at least 20× magnification or higher, and

can easily be more than $50,000 \times 50,000$ pixels, hence the term "gigapixel" images. These high-resolution images are in the range of $0.5-6$ GB on disk [10, 11]. Unlike other medical image modalities such as MRI and CT, common compression and storing techniques are not applicable for WSIs due to the image size and sensitive nature of anatomic clues in high magnification.

WSIs are stored in a layered pyramid structure, where each layer represents WSI at a different zoom (magnification) level [10]. This helps facilitate efficient and user-friendly WSI software tools, optimized for WSI visualization and analysis.

Figure 2.1 illustrates a WSI in three layers pyramid structure. The first layer, the base image, has the highest resolution, while the last layer, the thumbnail, has the lowest resolution [12]. Also, all the information about WSI such as resolution, size, and magnification is encoded in its header when formats like BigTIFF or SVS are used [13].

2.2 Machine learning in Computational Pathology

Digital pathology is becoming popular with the advent of whole-slide digital scanners resulting in growing availability of digitized histopathology images. Having access to WSIs not only makes pathologists' job easier and relieve the workload on pathologists [14] but also allows us to apply image analysis techniques in the histopathology field. Image analysis techniques include detection, segmentation, and classification, which can be beneficial for many clinical tasks. They can significantly reduce laborious and time-consuming tasks as they can provide us with robust quantifications, e.g., nuclei, mitosis, and cell counting. This quantitative analysis is important to understand the underlying reasons for diagnosis [15]. Additionally, sophisticated machine learning techniques can support medical centers and pathologists to gain a second opinions in cancer classification and detection tasks [16, 17]. Machine learning in computational pathology can be categorized into two groups: supervised and unsupervised learning. Supervised learning approaches rely on labeled data to infer a mapping function between inputs and labels. The main goal of unsupervised learning methods, in contrast, is to discover hidden structures in unlabeled data. However, applying both supervised and unsupervised techniques is not straightforward in histopathology and may require design of unique processing steps since WSIs have unique properties [18].

There are several challenges pertinent to applying machine learning techniques to high-resolution histopathology WSIs. However, machine learning offers many opportunities to develop efficient and robust methods, deal with those challenges, and improve diagnostic

interpretation in histopathology. Some of these challenges and opportunities are introduced in the following subsections.

2.2.1 Challenges

This section will focus on four challenges concerning representation and processing of WSIs through machine learning techniques, and identifying two opportunities of applying machine learning in histopathology [18, 19].

Large Dimensions

Many machine learning techniques have been applied to general datasets that contain natural images (animals, objects, etc.) of size 256×256 pixels. However, WSIs are extremely large images, which often exceed $50,000 \times 50,000$ pixels. These images are hard to analyze due to their massive size. It seems to be impossible to directly feed these images into a neural network as they can exhaust computational resources [19]. Also, Resizing WSIs to smaller sizes would lead to loss of critical information [18]. One common approach for this challenge is to divide WSIs into small regions called "patches", e.g., 1000×1000 patches. Then, each patch may be analyzed independently [16].

Multi-Magnification

Pathologists examine glass slides or WSIs by acquiring information from various magnification levels. They may obtain structural information at lower magnifications and detailed cellular information at higher magnifications. Although it is more common to use WSIs in 20x magnification for machine learning, it has been shown that combining images from different magnifications may improve the final performance of machine learning algorithms depending on diseases and tissue types [18, 20]. This thesis used WSIs at 20x magnification.

Lack of Diversified Datasets

The morphology of a single subtype of cancer can appear in different patterns. Morphology variability, or polymorphism, can make learning challenging as the model have to learn all possible patterns for each cancer type. Additionally, slide preparation, staining, and digtal scanners are different among different medical centers. As a result, training samples need to be gathered from multiple medical centers. This will enable the model to be more

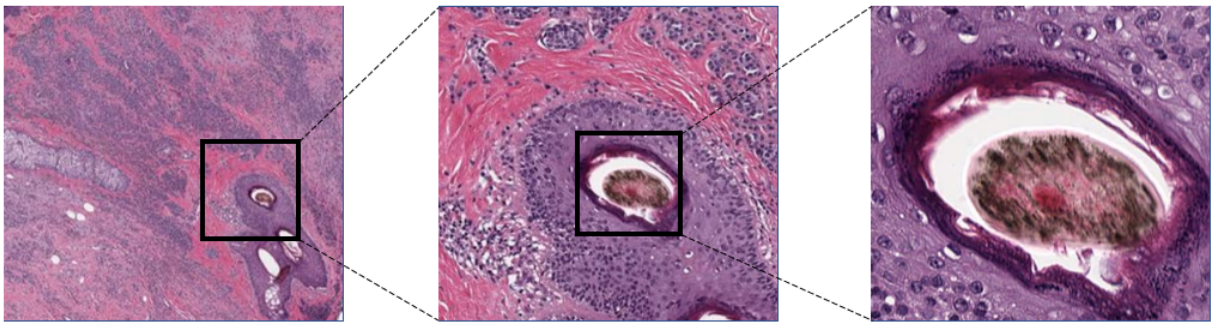


Figure 2.2: Patches of the same WSI at three different magnification levels. The left image at lower magnification has more structural tissue information while the right image at higher magnification has more detailed cellular information.

generalized by being trained with a diverse dataset [19]. In this regard, decentralized learning allows training the model on datasets from various organizations without the need to collect data in one central location.

Security and Privacy

Compared to data in other domains, healthcare data is highly sensitive. Medical records can reveal intimate details about patients' physical and mental health. Many guidelines and regulations stress the privacy, confidentiality, and security of patients, prohibiting medical centers from sharing data with other parties outside their organizations [21, 22]. At the same time, machine learning is inherently dependent on having data from diverse sources to train a generalized model. Obviously, federated learning as a distributed learning approach may be able to address this challenge. It allows multi-institutional collaboration on decentralized data, protecting the privacy of each individual participant.

2.2.2 Opportunities

Computer-Assisted Diagnosis

The main aspect of digital pathology is the computer-assisted diagnosis (CAD), which has the potential to revolutionize pathologists' routine tasks [18]. CAD can facilitate, quantitative assessment (e.g., nuclei and mitosis count), segmentation (e.g., contouring

structures and regions), classification (e.g., benign vs. malignant and cancer grading), and detection (e.g., detecting region of interest (ROI) such as tumor region). CAD not only relieves the workload on pathologists and saves their time to focus on complicated cases, but it could – when offered with well-trained models in a user-friendly fashion – provide clinics with a computerized second opinion which can significantly reduce inter- and intra-observer variability [15].

Distributed and Decentralized Learning

Recent advances in machine learning have opened new horizons in digital pathology. It has removed geographical barriers, helping medical centers to benefit from the diagnosed cases in other medical centers in a decentralized fashion. Distributed learning enables training a machine learning model without the need to collect data in one central location. It not only makes the process more agile and economic since it does not require centralized storage but also helps to maintain the privacy of the medical records as the data are never touched by off-site access in some decentralized learning methods. More details on distributed learning will be provided in the following section.

2.3 Distributed and Private Machine Learning

The prediction power of deep models depends, among others, on the size of the training data. Having access to large-scale datasets enables the model to more precisely estimate the underlying distribution of the data. For example, the public availability of the *ImageNet* dataset [23] with almost 1.2 million diverse data samples has led to considerable advancements in computer vision. Machine learning methods, and deep models, in particular, rely on training datasets that are ideally collected and aggregated from various sources. However, it appears currently impossible to gather large-scale datasets in one central location in medical domain due to privacy concerns [22]. In centralized learning methods for medical imaging, training data is from different medical centers such as hospitals and clinics, which are brought together to a central location, commonly called a *server*. However, hospitals are generally not willing to share their patients' medical records with external collaborators because of privacy considerations and regulatory compliance [24]. Therefore, the lack of large-scale diverse datasets publicly available to researchers hinders model development in healthcare. To overcome these challenges, decentralized learning methods appear to be a promising scheme to preserve data privacy while enabling training

of generalizable models. Federated methods allow training on multi-institutional datasets without requiring direct access to data.

Federated learning has emerged as a promising solution to protect user-sensitive data by keeping data local. It is a novel decentralized paradigm that plays a critical role in privacy-sensitive applications, opening a new horizon for secured decentralized learning methods. Federated learning as a decentralized approach was first proposed by McMahan in 2017 [25]. The main concern of federated learning is data privacy to protect users against data disclosures. It has become a recent active area of research because of increased concerns for data privacy and cybersecurity [26].

Federated learning was first proposed for wireless mobile network applications to accommodate decentralized training [27]. It enables mobile users to keep their local data private, training the model locally relying on their private data and sharing only the training results (not data) with the central server and hence other users. This allows the server to train the global model relying only on the aggregated training results of decentralized data. Although federated learning was first proposed for wireless networks, it can be applied to other privacy-sensitive applications such as the medical domain.

2.4 Federated Learning

There are three principal challenges in histopathology image analysis, which encourage us to employ federated learning. Firstly, histopathology image analysis is a complicated, time-consuming task since images are large with many details relevant for diagnosis. Using artificial-intelligent systems can reduce the workload on pathologists and facilitate diagnosis [15]. Secondly, studies have shown that observer variability in histopathology image diagnosis can be significant due to specific diagnostic biases such as the pathologist's initial training or working environment [2]. Access to a second (computational) opinion provided by machine learning can make pathologists more confident in their decision and complement their visual assessments. Thirdly, similar to other medical domains, histopathology images contain critical information that hospitals may not be able to share for the sake of patients' privacy [22]. Also, having access to histopathology data from different medical centers is a vital prerequisite to training a generalized model.

Federated learning seems to be a solution that can address all these challenges. It not only trains deep models to make the diagnosis task more reliable but also can decrease observer variability and improve consensus building by learning a collaborative model for quantification of tissue morphology without violating patients' privacy.

There are two main components in federated learning, namely a central server and some local centres [25]:

- A **local center** is a clinic or hospital that is willing to participate in the learning of a global deep model but is not interested in sharing any patient data with other organizations. The local center might also be called a participant or hospital in this thesis. Apparently, there exists more than one local center in any federated learning scenario.
- A **central server** is the main hub responsible for orchestrating decentralized learning. It facilitates the aggregation of training results from all the local centers and its main goal is to train a global model with acceptable performance for all participants. Therefore, it is also called an *aggregation center*.

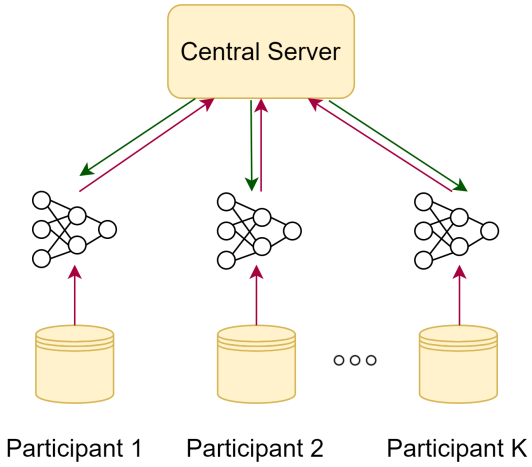


Figure 2.3: Illustration of federated learning with the central server and K participants.

Federated learning is a promising framework that addresses the fundamental challenges of privacy, ownership, and locality of data [26]. Depending on the task at hand, it enables training machine learning models, in a collaborative fashion, over decentralized data centers such as hospitals, relying on remote decentralized local data [25]. To learn a global model at the aggregation center, all local data centers periodically collaborate with each other through communication with the aggregation center. Figure 2.3 represents one possible architecture of a federated learning scheme with the central server and k participants. As

can be seen, each iteration of the federated learning procedure consists of three main steps described below [28]:

1. **Model distribution** – The aggregation center broadcasts the global model parameters to each of the local centers via a secured communication link.

2. **Local training** – Given the global model parameters at the previous stage, each local center trains the global model by relying on its local data, updating the model parameters locally.

3. **Global aggregation** – All participating local centers send the training results back to the aggregation server. Next, the central server processes the received updates from local centers by updating the global model parameters. This procedure is called *global training*.

These three steps are repeated until the global model training converges into a robust model. These steps are common among almost all federated learning methods [25]. More specifically, various federated learning methods may have different assumptions, formulations, and procedures for implementing each step of the general flow of decentralized learning.

2.4.1 Mathematical Formulation

In this section, mathematical notations and formulations are presented to explain the formalism of federated learning [25, 29]. It is assumed that there is a fixed number of m participant hospitals $\{\mathcal{P}_i\}_{i=1}^m$ in our setting. Each of these participant hospitals has a fixed local dataset, denoted by $\{\mathcal{D}_i\}_{i=1}^m$ where each has n_i training samples. These m hospitals wish to train a model with parameter w. In conventional approaches, all data $\{\mathcal{D}_i\}_{i=1}^m$ is collected at one central location to train a machine learning model where data owners expose their data to other parties outside their hospitals. However, in federated learning data owners collaboratively train a model with parameters w without sharing their private data $\{\mathcal{D}_i\}_{i=1}^m$. It is assumed that all the participant hospitals take part in each round of training. In a typical machine learning problem, model weights are adjusted using the optimization problem

$$\min_{w \in \mathbb{R}^d} f(w),$$

where

$$f(w) \stackrel{def}{=} \frac{1}{n} \sum_{i=1}^{n} f_i(w),$$

and n is the number of samples in the training dataset. Function $f(w)$ is the loss function, and $f_i(w)$ is defined by

$$f_i(w) = \ell(x_i, y_i; w),$$

which is the loss of the prediction on training sample (x_i, y_i) given the model parameter w.

Assuming that the dataset is partitioned into m smaller datasets, the optimization problem above can be rewritten as follow

$$\min_{w \in \mathbb{R}^d} f(w) = \sum_{k=1}^{m} F_k(w),$$

where

$$F_k(w) = \frac{1}{n_k} \sum_{i \in \mathcal{P}_k} f_i(w).$$

In the optimization problem above, it is assumed that all those m datasets are in one central location and accessible for training. However, federated learning collaboratively trains a model without the need to collect and access private training samples. In the following section, two common federated learning benchmarks will be discussed in detail.

2.4.2 Benchmarking Federated Learning Methods

There are two common benchmark methods in federated learning, namely FedSGD (Federated Stochastic Gradient Descent) and FedAvg (Federated Averaging) [25]. FedSGD relies on stochastic gradient descent (SGD) for optimization, and FedAvg is built on top of FedSGD.

FedSGD

Federated stochastic gradient descent, FedSGD, is one of the benchmarks for federated learning [25]. In FedSGD, participants perform training only for one batch of data, meaning that one single gradient calculation is done per round of communication. The central server is the main node that can update model parameters by collecting gradients from

all participants. This approach is computationally efficient as hospitals calculate only one single batch gradient. However, it is expensive in terms of the number of communications to converge.

Given the learning rate η, each client k compute single batch one step of gradient $\Delta F_k(w_t)$ at current model w_t, creating a new local model weight w_{t+1}^k as follow

$$w_{t+1}^k \leftarrow w_t - \eta \Delta F_k(w_t).$$

Then, the central server aggregates local weights, updating the model parameters as follows

$$w_{t+1} \leftarrow \sum_{k=1}^{m} \frac{n_k}{n} w_{t+1}^k,$$

where n_k is the number of samples in the kth hospital, and n is the total number of training samples from all the participant hospitals. More detail on the FedSGD method is provided in Algorithm 1.

Algorithm 1 FedSGD [25]. There are m local centers, T is the number of epochs, η is learning rate, n_k is # samples in kth local center, n is total # samples, and w is the global model parameters.

Input: m, T, w^0, η
Output: w^{T-1}

1: **for** $t = 0, \ldots, T-1$ **do**
2: Server sends w^t to all hospitals
3: **for** $k = 1, 2, \ldots, m$ **do**
3: Given (k, w^t, η), each hospital updates the model for one batch of its data and returns w_k^{t+1}
4: **end for**
5: Each hospital k sends w_k^t back to the server
6: Server updates w^{t+1} as

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

7: **end for**
8: **return** w^{T-1}

Algorithm 2 FedAvg [25]. There are m local centers, T is the number of epochs, η is learning rate, n_k is # samples in kth local center, n is total # samples, B is local minibatch size, E is the number of local epochs, and w is the global model parameters.

Input: m, C, T, w^0, η, n_c
Output: w^{T-1}

for $t = 0, \ldots, T-1$ **do**
  Server sends w^t to all hospitals
  % *Local Training*
  for $k = 1, 2, \ldots, m$ **do**
    $w_k^{t+1} \leftarrow$ **LocalTraining**(k, w^t, η) % update weights
  end for
  % *Global Training*
  Server updates w^{t+1} as

$$w^{t+1} \leftarrow \sum_{k=1}^{m} \frac{n_k}{n} w_k^{t+1}$$

end for
return w^{T-1}

LocalTraining(i, w_t, η) :

$\mathcal{B} \leftarrow$ (split dataset of ith hospital into batches of size B)
for local epoch $j = 1, 2, \ldots, E$ **do**
  for batch $b \in \mathcal{B}$ **do**
    $w \leftarrow w_t - \eta \nabla F_k(w_t; b)$ % $F_k(.)$ *is the loss function for hospital* k
  end for
end for
return w

FedAvg

Another common benchmark method for federated learning is FedAvg (federated averaging). FedAvg is the expansion of FedSGD, where more computation is added to each client by iterating the local updates. In FedAvg [25], each participating hospital trains the global model on its local data for some number of batches and epochs. Then, the central server aggregates the training results and simply takes the average of training results across all hospitals. FedAvg algorithm has been provided in detail in Algorithm 2.

Comparing FedAvg and FedSGD

Table 2.1 summarises the advantages and disadvantages of FedSGD and FedAvg approaches. Firstly, in FedSGD, hospitals need to send their training results to the central server for every batch of their local data, which leads to high communication overhead. However, in FedAvg, hospitals update model parameters for multiple batches and epochs, then send training results to the central server. Therefore, FedAvg has moderate communication overhead compared to FedSGD at the expense of having computational cost in each hospital. Secondly, FedAvg allows training to be faster compared to FedSGD since both hospitals and central server can update the model in FedAvg. Thirdly, in FedSGD, the central server has access to accurate gradient information and updates the model based on those gradient, therefore, the convergence is guaranteed in this approach. However, in FedAvg, the central server does not have access to model gradients. It just takes average over local updates, therefore, model convergence is not guaranteed in this approach.

Table 2.1: Performance Comparison between two federated learning benchmarks, FedSGD and FedAVG [29].

Method	**Advantages**	**Disadvantages**
FedSGD	Accurate gradient information Guaranteed convergence Computationally efficient	High communication cost Slow convergence Require reliable connection
FedAvg	Faster convergence Communication efficient	High computational cost at hospitals No guarantee of convergence

2.4.3 Federated Learning in Medical Domain

Strict national and regional privacy rules, such as General Data Protection Regulation in Europe or HIPAA in the United States, prohibit medical centers to share patients' medical records [30] . However, conventional machine learning methods require to access data in one central location. To mitigate these challenges, the research community has recently started to explore the potentials of federated learning for medical image analysis, creating multi-center healthcare ecosystem. Federated learning has the potential to increase both the size and diversity of training data without violating data privacy. It enables medical centers to deploy large-scale machine learning models trained on different data centers without sharing sensitive private data. Federated learning has been employed for classification,

segmentation, and detection tasks applied on various types of medical images, such as CT, MRI, microscopy images, and WSI patches.

For detection examples, authors in [31] used federated learning for abnormal detection in COVID 19 CT images, and authors in [32] used federated learning for abnormal tissue detection in brain MRI images. Segmentation of tumors in brain MRI images in federated learning setting has been studied in [33]. The authors in [34] used private prostate MRI images from three institutions and showed the superiority of federated learning in prostate MRI segmentation. As another example of segmentation, authors in [35] employed federated learning for segmenting COVID region in chest CT images. As classification examples, skin lesion classification in decentralized scenario has been studied in [36]. Authors in [37] conducted a case study of federated learning on histopathology images with the focus on differential privacy. They have also examined design factors on the classifier trained in federated learning setting. Classification of histopathology images in federated learning fashion have also been investigated in [38]. The main goal of the authors in [38] is to reduce the amount of communication between the server and hospitals.

These papers mostly focus on studying the applicability of federated learning in the medical domain. They do not address urgent challenges in applying federated learning for medical images.

2.5 Summary

This chapter introduced the necessary definitions and concepts that readers should be familiar with to understand the content presented in upcoming1 chapters. Recent advances in digital pathology by enabling whole slide imaging are acting as the key enablers of machine learning in computational pathology. Federated learning was introduced as a promising remedy for medical data privacy concerns. Most existing works on federated learning in the medical domain involve applying existing federated learning approaches on medical data. They do not address the real-world challenges of federated learning methods on medical data.

In the next two chapters of this thesis, two real-world challenges of federated learning will be addressed. Secure communication between hospitals and the central server and fairness in aggregation of local training results will be discussed in Chapter 3 and Chapter 4, respectively.

There are some assumptions in this thesis listed as follows

- Data has acceptable quality such that the machine learning models can be trained based on their content.
- Local data in each hospital is reliable and there is no malicious hospital willing to attack learning.
- Hospitals are willing to collaborate to train a model.
- The communication link between hospitals and the central server is stable.
- In each round of training, hospitals can be synced with the central server.

The majority of the existing works in the literature implicitly consider these assumptions without explicitly mentioning them.

Chapter 3

Privacy in Federated Learning

Federated learning is a novel machine learning method enabling hospitals to collaboratively learn a model without sharing private patient data for training. In federated learning, participant hospitals periodically exchange training results rather than training samples with a central server or other hospitals. However, having access to model parameters or gradients can expose private training data samples. To address this challenge, this thesis proposes to adopt secure multiparty computation (SMC) to establish a privacy-preserving federated learning framework.

Two different SMC-based privacy-preserving framework are proposed for two different federated learning architectures, i.e., centralized and decentralized federated learning. These two architectures will be discussed in the following sections.

Experiments are conducted on a publicly available repository, The Cancer Genome Atlas (TCGA) [39]. The performance of the proposed framework was compared with differential privacy and federated averaging as the baseline. The results reveal that compared to differential privacy, the proposed framework can achieve higher accuracy with no privacy leakage risk at a cost of higher communication overhead.

3.1 Introduction

Federated learning enables learning a model while all participants keep data private, sharing training results with the central server. However, authors in [40] have shown that sharing the model's parameters or gradients is not safe. They demonstrate that having access to the model's weights or gradients can expose training samples. Therefore, privacy-preserving

methods in federated learning have recently been introduced to protect training samples from leakage.

3.1.1 Privacy Preserving Methods

There are three different strategies for privacy-preserving federated learning in the literature to securely share the training results [41, 24].

Differential Privacy (DP)

Differential Privacy(DP) was originally developed to enhance secure analysis over sensitive data. DP protects privacy by adding noise to the training results before sharing with the central server [42]. Although perturbing the training results improves the privacy of the training samples, it might adversely impact accuracy of the model.

Secure Multiparty Computation (SMC)

Secure Multiparty Computation (SMC) was first introduced as a secure two-party computation, then generalized in 1986 [43]. SMC allows us to compute functions of private input values such that each party learns only the corresponding function output value, but not input values from other parties. Now, SMC [44] is a privacy-preserving method, enabling hospitals to jointly compute a function on their model's weight without revealing the actual weights values. Unlike DP, SMC does not perturb the training results. However, compared to DP, it has communication overhead since hospitals communicate with each other to compute the weights average.

Homomorphic Encryption (HE)

The idea of Homomorphic Encryption (HE) was first introduced in 1978 as a solution to perform computation over ciphertext without decrypting the ciphertext [45]. HE relies on encoding/decoding gradients and uses encrypted data for training [46]. It allows computation on encrypted gradients and decryption of the results is equivalent to performing the same operations on gradients without any encryption. This method is efficient in terms of communication cost, however, it is computationally expensive.

The effectiveness of DP in decentralized learning has been investigated in the healthcare domain [47, 37]. Authors in [47] preserve accuracy by adding Gaussian noise to the trained

model weights, providing extensive experiments on MRI images. In [37], the authors conduct the feasibility study of DP in federated learning. Also, the impact of the design factors of DP in decentralized learning has been investigated on histopathology images.

SMC has played a successful role in cloud computing and the Internet of Things (IoT) [48]. Recently, SMC has been adopted as a privacy-preserving method in federated learning. For example, authors in [49] applied chained SMC in FL to protect model weights from disclosure. In their framework, first, the central server sends one of the participants a random number. Then participants sequentially communicate with each other to compute the average of the local models. This framework imposes extreme latency and cannot be scaled since all the participant has to communicate sequentially. However, this chapter of the thesis will propose a framework such that communications happen in parallel within clusters. The proposed methods address the privacy challenges of federated learning by introducing novel frameworks based on SMC. Unlike DP, SMC does not compromise the model accuracy since it does not perturb training results.

3.2 Federated Learning Architecture

There are two common architectures for federated learning, namely centralized and decentralized federated learning. The key difference between these two groups is about the existence of the central server. In this section, these two popular architectures will be described [29].

3.2.1 Centralized Federated learning Architecture

A typical centralized architecture of federated learning is shown in Figure 3.1, which is also called client-server architecture. In this system, participant hospitals rely on the central server and collaboratively learn a machine learning model. In this architecture, hospitals locally train the model, and the central server is responsible for the aggregation of local training results.

3.2.2 Decentralized Federated learning Architecture

In addition to the client-server introduced above, another common architecture is decentralized federated learning which is shown in Figure 3.1. This architecture is also called

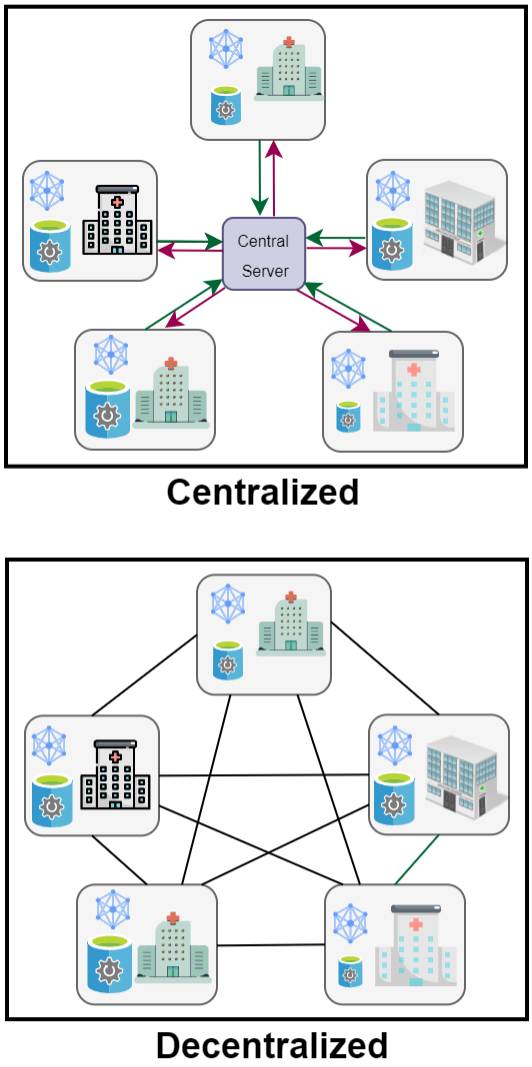


Figure 3.1: Two types of federated learning from networking structure perspective: centralized versus decentralized federated learning.

peer-to-peer. Under this architecture, there is no central server or co-ordinator. As shown in Figure 3.1, in this architecture, first participants train the model on their local data. Then, they transfer their training results to each other. Finally, each participant updates the model based on all the received training results from other hospitals. As there is no central server to aggregate the training results, participants need to agree on the protocols for sending and receiving training results in advance.

3.3 Privacy in Centralized Federated Learning

In this section, an SMC-based framework for centralized federated learning architecture will be introduced. In the proposed method, hospitals are divided into "clusters". Each hospital performs model training on its own local data. After local training, each hospital splits its model weights among other hospitals in the same cluster such that no single hospital can retrieve other hospitals' weights on its own. Then, all hospitals sum up the received weights, sending the results to the central server. Finally, the central server aggregates the results, retrieving the average of the models' weights and updating the model without having access to individual hospitals' weights.

3.3.1 Method

In this section, the proposed SMC-based FL method will be introduced in detail. Figure 3.2 represents the proposed cluster-based SMC framework for centralized federated learning. There are K hospitals, which will be equally divided into M clusters with size $N = K/M$. Each hospital belongs to one cluster which is denoted by $c = \{1, \ldots, M\}$. Hospital k in cluster c is represented by H_k^c. The set n_c with length N represents indexes of all hospitals in cluster c. Model training in the proposed approach is performed in three steps, local training, SMC, and aggregation.

Step 1: Local Training– All participant hospitals train the model individually with their local data, updating the model. Model parameters trained by the kth hospital is denoted by w_k.

Step 2: SMC– Hospital H_k^c generates N random numbers $\{\beta_{k,j}^c | 0 < \beta_{k,j}^c < 1, j \in n_c\}$ that sum up to one.

$$\sum_{j \in n_c} \beta_{k,j}^c = 1. \tag{3.1}$$

Algorithm 3 Proposed method for **centralized** federated learning. There are K hospitals, M clusters, T is the number of epochs, E is the number of local epochs, η is learning rate, n_c index of all hospitals in cluster c.

Input: M, C, T, w^0, η, n_c
Output: w^{T-1}

for $t = 0, \ldots, T-1$ **do**
  Server sends w^t to all hospitals
  % *Step1: Local Training*
  for $k = 1, 2, \ldots, K$ **do**
    $w_k^{t+1} \leftarrow$ **LocalTraining**(k, w^t, η) % update weights
  end for
  % *SMC*
  $R_k^c \leftarrow 0$
  for $c = 1, 2, \ldots, M$ **do**
    for $k \in n_c$ **do**
      for $i \in n_c$ **do**
        $R_k^c += \beta_{i,k}^c w_i^t$
      end for
      Hospital k feedbacks R_k^c to the central server.
    end for
  end for
  % *Step3: Aggregation*
  Server updates w^{t+1} as

$$w^{t+1} \leftarrow \frac{1}{K} \sum_{k=1}^{K} R_k^c$$

end for
return w^{T-1}

LocalTraining(i, w_t, η) :

$\mathcal{B} \leftarrow$ (split dataset of ith hospital into batches of size B)
for local epoch $j = 1, 2, \ldots, E$ **do**
  for batch $b \in \mathcal{B}$ **do**
    $w \leftarrow w_t - \eta \nabla F_k(w_t; b)$ % $F_k(.)$ *is the loss function for hospital* k
  end for
end for
return w

Then, each hospital k in cluster c, H_k^c, sends portions of its own locally trained model parameters to each of $N-1$ neighbours in cluster c. Mathematically, H_k^c sends $\beta_{k,j}^c w_k$ to hospital j for all $j \in n_c$. In the end, the kth hospital will have some portion of its own, and some portion of its $N-1$ neighbor's model parameters as follows:

$$\mathcal{H}_k^c : R_k^c = \sum_{i \in n_c} \beta_{i,k}^c w_i. \tag{3.2}$$

Step 3: Aggregation– Finally, each hospital sends R_k^c to the central server, and the central server takes the average of R_k^c of all the hospitals in all clusters as follows:

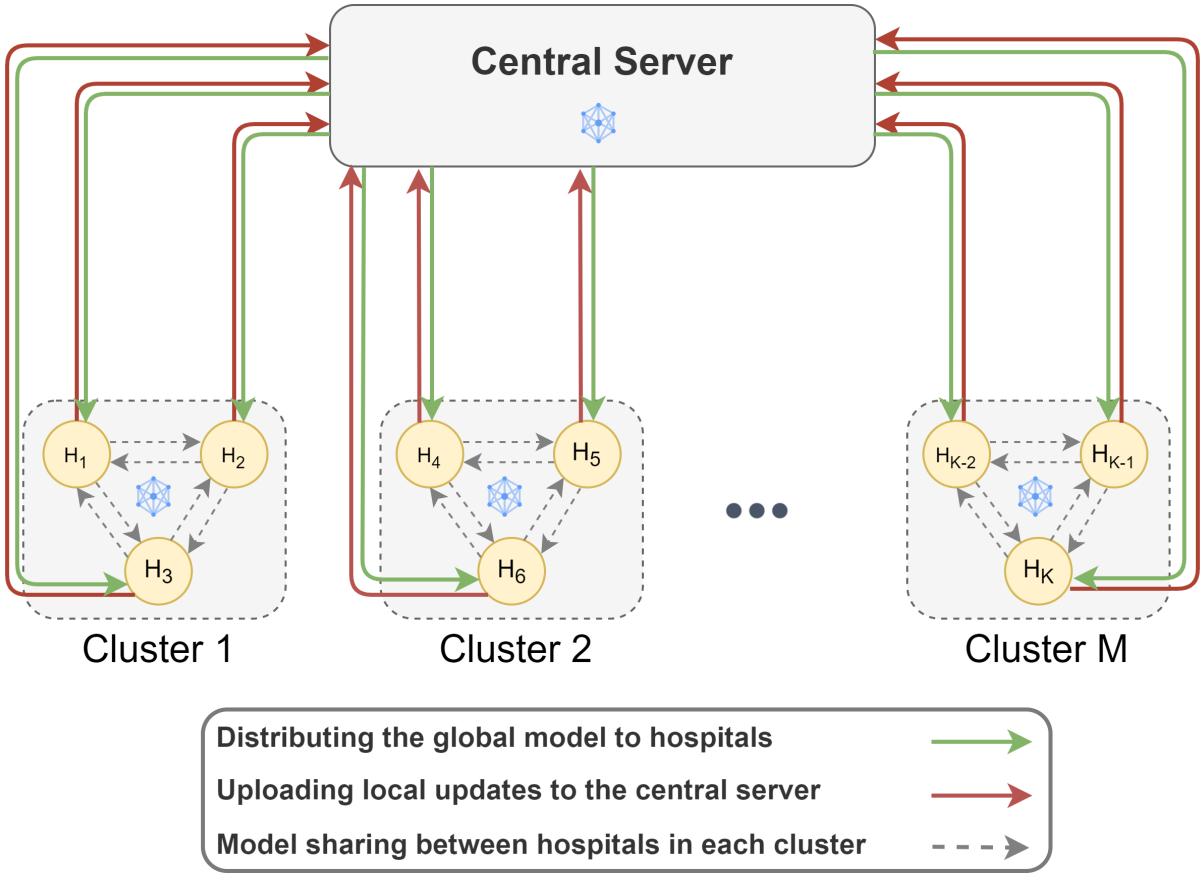


Figure 3.2: Cluster-based SMC for **centralized** federated learning.

Table 3.1: The summary of the dataset [37].

Client	**# Slides**	**# Patches**
C1: Int. Gen. Cons.	267	66,483
C2: Indivumed	211	52,539
C3: Asterand	207	51,543
C4: Johns Hopkins	199	49,551
C5: Christiana H.	223	55,527
C6: Roswell Park	110	27,390

$$
\begin{aligned}
w &= \frac{1}{K} \sum_{c=1}^{M} \sum_{k \in n_c} R_k^c \\
&= \frac{1}{K} \sum_{c=1}^{M} \sum_{k \in n_c} \sum_{i \in n_c} \beta_{i,k}^c w_i.
\end{aligned}
\tag{3.3}
$$

If we exchange the position of the two summations in Eq. 3.3, we will get

$$
\begin{aligned}
w &= \frac{1}{K} \sum_{c=1}^{M} \sum_{i \in n_c} \underbrace{\sum_{k \in n_c} \beta_{i,k}^c}_{1} w_i \\
&= \frac{1}{K} \sum_{c=1}^{M} \sum_{i \in n_c} w_i \\
&= \frac{1}{K} \sum_{i=1}^{K} w_i.
\end{aligned}
$$

As shown above, the central server can recover the exact average of local weights without having access to the weights of each individual hospital. These steps have been summarized in Algorithm 3.

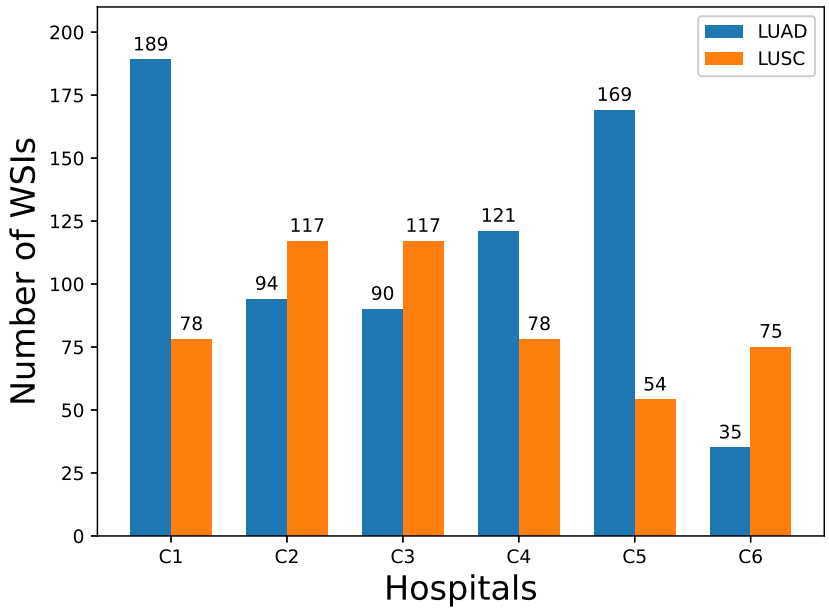


Figure 3.3: Label distribution in dataset.

3.3.2 Experiments and Results

Datasets

The proposed privacy-preserving federated learning is evaluated on The Cancer Genome Atlas (TCGA) [39, 50] dataset, the largest publicly available archive of histopathology WSIs. This dataset has more than $30,000$ H&E stained WSIs that have been collected from various medical centers all over the world. To validate the proposed method, TCGA WSIs diagnosed with non-small cell lung cancer (NSCLC) were selected to construct a dataset of multiple institutions. This cancer has two frequent subtypes, namely

- Lung Adenocarcinoma (LUAD)
- Lung Squamous Cell Carcinoma (LUSC).

This study includes hospitals that have WSIs from both LUAD and LUSC subtypes. In TCGA, only six hospitals met this requirement. Therefore, WSIs diagnosed with NSCLC were collected from those six hospitals to create the dataset with six participants. As

mentioned before, WSIs are extremely large files. Therefore, they cannot directly be fed into any neural network. The common approach to deal with these images is to divide them into patches of smaller sizes for further analysis [51]. The selected WSIs were divided into patches of size 1000×1000 pixels. In [37] more details are provided on patch extraction and selection of the lung dataset that has been used in the experiments. The statistics of this dataset for each hospital are presented in Table 3.1 and Figure 3.3. The dataset of each hospital has been randomly divided into 80% and 20% groups for training and testing purposes, respectively.

Experimental Details

Figure 3.4 illustrates WSI pre-processing as well as the model used to classify lung samples into LUAD and LUSC subtypes. As shown in this figure, for the classification of lung histopathology WSIs, pre-trained DenseNet121 [52] was employed to extract features of length 1024 for each patch. The pre-trained weights on the ImageNet dataset for DenseNet121 have been used with no additional adjustment. Next, attention-gated multiple instance learning (MIL) [53] was used to combine feature vectors of patches of each WSI, creating a feature of size 1024 for each WSI classification [53]. The reason for using MIL is that when a WSI is divided into multiple patches, we are dealing with instances for which only a single WSI level label, namely a primary diagnosis, is provided. Therefore, MIL architecture is required to learn a model that can predict the WSI label given a bag of instances (patches). The attention-based MIL architecture enables the model to combine the features of patches to create one feature vector of length 1024 that will be used for the classification of WSI. This architecture aggregates feature vectors of those patches such that key patches are assigned relatively higher weights. The high-level structure of the MIL classifier has been visualized in Figure 3.4. The MIL gated attention classifier is the network that is learned in a decentralized federated learning fashion. More details on this MIL network is provided in [53].

Results and Discussions

In this section, the experimental results are presented for the lung histopathology dataset. The proposed SMC based method is compared with the baseline which is FedAVG [25] without any privacy-preserving consideration. Also, the proposed method will be compared with DP which has been implemented on top of FedAvg. An ideal privacy-preserving method have to have a closed performance to the baseline while preserving privacy of training results. The histopathology lung dataset includes data from six hospitals. Those

$K = 6$ hospitals are divided into $M = 2$ clusters of size $N = 3$. DP is deployed according to [47] with additive Gaussian noise standard deviation of 0.03. The standard deviation has been selected to have the highest possible privacy while the classification performance is still acceptable. For all these three methods, an Adam optimizer was used with the

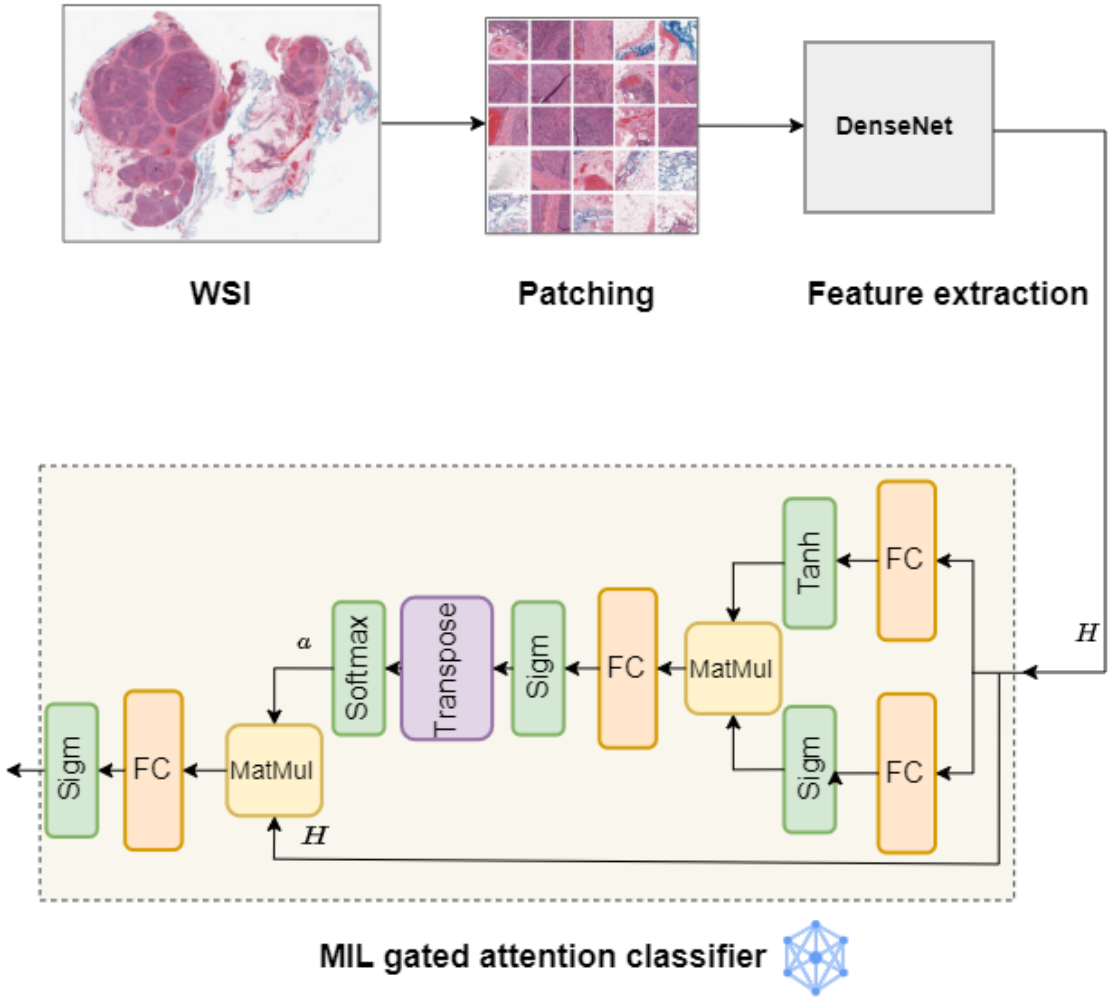


Figure 3.4: The illustration of the end-to-end training procedure. First, WSIs are divided into patches of size 1000×1000. Next, the features of patches are extracted using DensNet121. Finally, those features are fed into a MIL gated attention classifier.

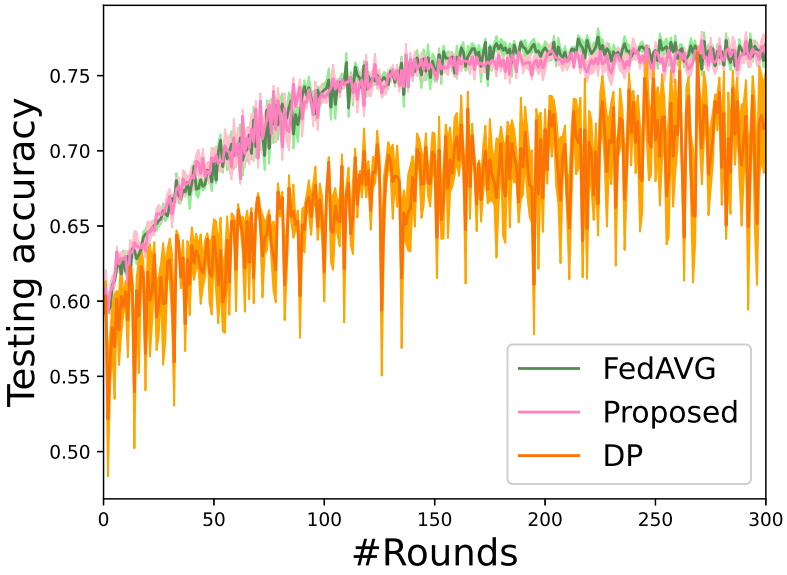


Figure 3.5: The average testing accuracy for 300 rounds of training over all hospitals.

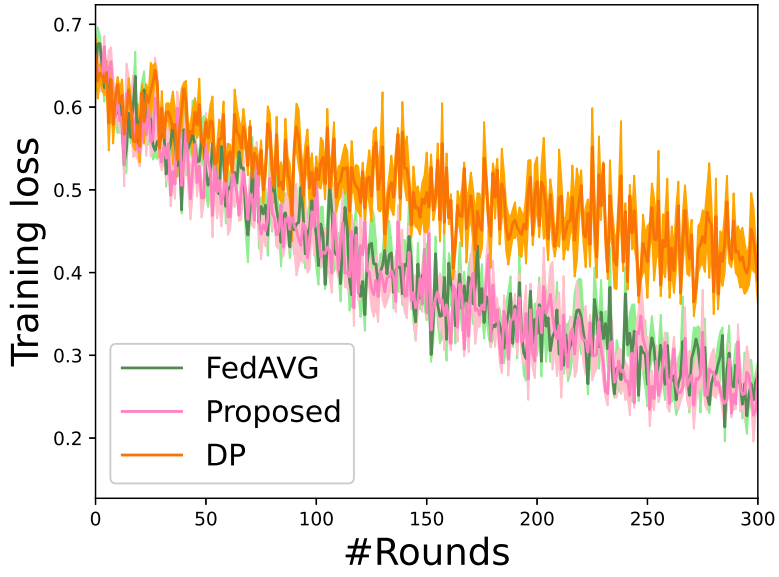


Figure 3.6: The average training loss and testing accuracy for 300 rounds of training over all hospitals for the proposed SMC framework in centralized federated learning.

following setting: epochs=300, batch size=32, number of local epochs=1, and learning rate=0.009.

Table 3.2 shows the performance of each method for each hospital in terms of accuracy and F1 Score. As represented, in each hospital, the proposed method has a closed performance to the baseline and outperforms DP. Additionally, the average performance of the proposed method surpasses DP. Figure 3.5 and 3.6 compare methods in terms of the average testing accuracy and average training loss of participant hospitals for 300 rounds of training communication between hospitals and the central server. As can be seen, the

Table 3.2: Experimental results of the proposed SMC based method for **centralized** federated learning.

Client	**Method**	**Accuracy**	**F1-Score**
C1: Int. Gen. Cons.	FedAvg	76.38	82.51
	DP	66.12	69.89
	Proposed	75.01	81.08
C2: Indivumed	FedAvg	85.46	87.63
	DP	79.06	81.12
	Proposed	87.20	89.03
C3: Asterand	FedAvg	81.54	80.96
	DP	74.40	70.27
	Proposed	80.95	80.47
C4: Johns Hopkins	FedAvg	75.01	82.74
	DP	69.37	73.84
	Proposed	75.62	83.12
C5: Christiana H.	FedAvg	73.33	82.31
	DP	64.87	68.54
	Proposed	68.88	78.58
C6: Rosewell Park	FedAvg	68.18	66.74
	DP	68.18	63.34
	Proposed	69.31	66.78
Avgerage	FedAvg	76.65	80.48
	DP	70.33	71.16
	Proposed	76.16	79.84

proposed method performs close to the baseline, surpassing DP. To eliminate the impact of random parameters in the experiments, all experiments has been repeated five times and all the results have been provided by taking the average over these five runs.

3.4 Privacy in Decentralized Federated Learning

Decentralized Federated learning is a privacy-preserving machine learning approach that allows training in a collaborative fashion. It enables training the model on data from various medical centers without any central point that coordinates the training process between hospitals. Decentralized federated learning maintains data privacy by allowing each hospital to train the model on its local data and exchange the training results rather than training samples. However, it has been shown in the literature that training results are as important as training samples since they can expose training samples. This challenge will be addressed in decentralized federated learning by employing secure multi-party computation. The proposed framework consists of three phases, and the first two phases occur only one time at the beginning of the training procedure. The main purpose of Phase I and Phase II is to learn the summation of the noise that each hospital uses to protect its training results from revealing. In Phase III, the participant hospitals train the model on their local data, adding noise to the training results and sharing the training results with other hospitals. Then, each hospital aggregates the training results and removes cumulative noise relying on Phase I and Phase II, and updates the global model. In Phase III, training process continues until convergence happens or reaching the maximum number of training rounds.

3.4.1 Method

In this section, details of the proposed SMC-based framework for decentralized federated learning will be introduced. It is assumed that there are K hospitals willing to participate in training the global model in a decentralized fashion. The ith hospital is denoted with H_i. Model training in the proposed method is performed in three phases. Figures 3.7, 3.8, and 3.9 represent those phases for $K = 3$.

- **Phase I** – As shown in Figure 3.7, H_1 generates two random numbers, N_1 and α, and send their summation to the next hospital, H_2. Then, H_2 generates a random number N_2, adds it to what has been received from H_1, and sends $N_1 + N_2 + \alpha$ to the next hospital H_3. Finally, H_3 generates random number N_3, add it to its received

number, return $\sum_{k=1}^{K} N_k + \alpha$ to the first hospital. Since the first hospitals knows the amount α, it can recover $\sum_{k=1}^{K} N_k$.

- **Phase II** – This phase is illustrated in Figure 3.8. As shown, the first hospital shares $\sum_{k=1}^{K} N_k$ with other hospitals. Therefore, each hospital knows the accumulative of random numbers generated by all hospitals.

- **Phase III** – Figure 3.9 represents Phase III of the proposed framework. In this phase, each hospital k performs local training, adds N_k to its training results, and shares it with all other hospitals. Denoting the training results of the kth hospital with G_k, each hospital will have $\sum_{k=1}^{K} G_k + \sum_{k=1}^{K} N_k$. From Phase I and Phase II, hospitals know $\sum_{k=1}^{K} N_k$, therefore, they can recover $\sum_{k=1}^{K} G_k$ and find the model average. This process in Phase III is repeated until convergence or reaching the maximum number of training rounds.

Phase I and Phase II take place only one time at the beginning of the learning procedure. These two phases help the participant hospitals to learn $\sum_{i=1}^{K} N_i$ which will be used later in Phase III. More detail on the proposed SMC-based approach for decentralized federated learning is provided in Algorithm 4.

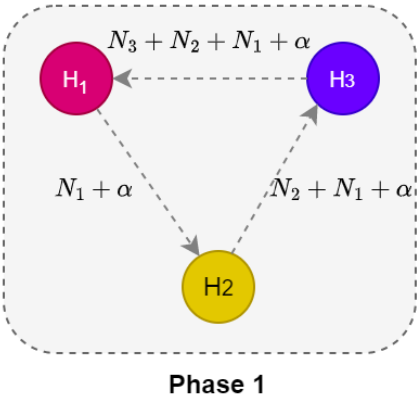


Figure 3.7: Phase I of the proposed SMC-based framework for decentralized federated learning. This phase happens only one time at the beginning of the training.

Algorithm 4 Proposed method for **decentralized** federated learning. There are K hospitals, T is the number of epochs, E is the number of local epochs, η is learning rate.

Input: K, T, w^0, η
Output: w^{T-1}

Hospital 1 generates random number N_1 and α and send $N_1 + \alpha$ to hospital 2. % *Phase I*
for j=2, . . . , K - 1 **do**
    Hospital j generates random number N_i
    Hospital j sends $\sum_{i=1}^{j} N_i + \alpha$ to hospital $j + 1$
end for
Hospital K returns $\sum_{i=1}^{K} N_i + \alpha$ to hospital 1
Hospital 1 recovers $\sum_{i=1}^{K} N_i$
% *Phase II*
for i = 2, . . . , M **do**
    Hospital 1 sends $\sum_{i=1}^{K} N_i$ to hospital i
end for
% *Phase III*
for $t = 0, \ldots, T-1$ **do**
    for $k = 1, 2, \ldots, K$ **do**
        $w_k^{t+1} \leftarrow$ **LocalTraining**(k, w^t, η) % local updates
    end for
    for $k = 1, \ldots, K$ **do**
        Hospital k sends $w_k^{t+1} + N_k$ to all other hospitals
    end for
    At the end, hospital k receives $R_k^{t+1} = \sum_{k=1}^{K} w_k^{t+1} + \sum_{k=1}^{K} N_k$
    % *Aggregation*
    for $i = 1, \ldots, K$ **do**
        Hospital i computes $\sum_{k=1}^{K} w_k^{t+1} = R_k^{t+1} - \sum_{k=1}^{K} N_k$
        $w^{t+1} \leftarrow \frac{1}{K} \sum_{k=1}^{K} w_k^{t+1}$
    end for
end for
return w^{T-1}

LocalTraining(i, w_t, η) :

$\mathcal{B} \leftarrow$ (split dataset of ith hospital into batches of size B)
for local epoch $j = 1, 2, \ldots, E$ **do**
    for batch $b \in \mathcal{B}$ **do**
        $w \leftarrow w_t - \eta \nabla F_k(w_t; b)$ % $F_k(.)$ *is the loss function for hospital* k
    end for
end for
return w

3.4.2 Experiments and Results

The proposed privacy-preserving frameworks are evaluated on TCGA data. WSIs diagnosed with non-small cell lung cancer (NSCLC) are selected to construct a dataset of six

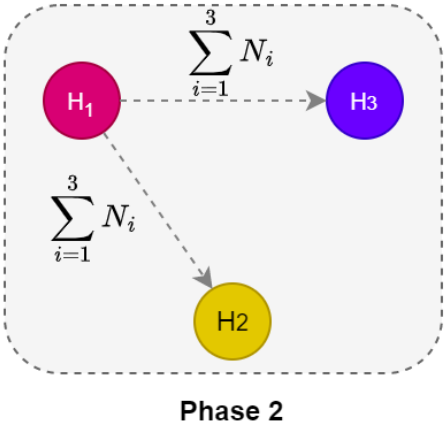


Figure 3.8: Phase II of the proposed SMC-based framework for decentralized federated learning. This phase happens only one time at the beginning of the training.

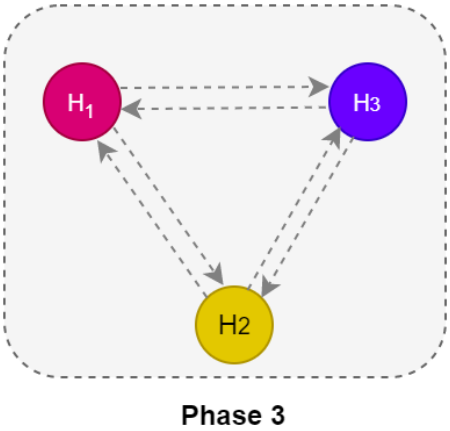


Figure 3.9: Phase III of the proposed SMC-based framework for decentralized federated learning. Unlike previous phases, Phase III runs multiple times to train the global model.

institutions. This cancer has two sub-types, Lung Adenocarcinoma (LUAD) and Lung Squamous Cell Carcinoma (LUSC). The dataset and classifier model that is used for this experiment is the same as in section 3.3.2. The classifier is trained to predict NSCLC sub-types. Table. 3.3 represents model performance in terms of accuracy and F1-Score. The proposed method outperforms DP while having a similar performance to the baseline.

Table 3.3: Results of the proposed SMC method for **decentralized** federated learning.

Client	**Method**	**Accuracy**	**F1-Score**
C1: Int. Gen. Cons.	FedAvg	76.91	82.06
	DP	67.95	70.42
	Proposed	76.82	81.81
C2: Indivumed	FedAvg	86.84	87.36
	DP	81.85	82.64
	Proposed	88.01	88.92
C3: Asterand	FedAvg	81.29	80.19
	DP	74.95	71.49
	Proposed	80.12	79.85
C4: Johns Hopkins	FedAvg	75.14	82.80
	DP	71.46	74.95
	Proposed	75.93	83.46
C5: Christiana H.	FedAvg	71.84	81.53
	DP	66.75	70.32
	Proposed	69.21	79.54
C6: Rosewell Park	FedAvg	68.93	68.72
	DP	68.13	64.97
	Proposed	69.94	67.23
Average	FedAvg	76.82	80.44
	DP	71.84	72.46
	Proposed	76.67	80.13

3.5 Summary

This section addressed the privacy-preserving challenge of federated learning. Two SMC-based frameworks were proposed to protect individual hospitals' model parameters from disclosure in both centralized and decentralized federated learning. In the proposed methods, neither participant hospitals nor the central server has access to the model weights of individual hospitals; however, the weights average can be recovered in the aggregation phase. The experimental results suggested that the proposed methods outperform DP in terms of accuracy and F1 Score at the expense of more communication overhead. However, having slight communication overhead to get higher accuracy is most likely acceptable in the medical domain. Additionally, each hospital needs to perform pre-processing to find suitable additive noise standard deviation in the DP method. However, the proposed methods do not require any pre-processing since they do not have any hyperparameters. Therefore, depending on the application, applying SMC-based frameworks for privacy-preserving purposes might be preferable compared to other privacy-preserving methods such as DP.

Chapter 4

Fair Federated Learning

Medical centers and healthcare providers have concerns and hence restrictions around sharing data with external collaborators outside their organizations. Federated learning, as a privacy-preserving method, involves learning a site-independent model without having direct access to patient-sensitive data in a distributed collaborative fashion. The federated approach relies on decentralized data distribution from various hospitals and clinics. The collaboratively learned global model is supposed to have acceptable performance for the individual sites. However, existing methods focus on minimizing the average of the aggregated loss functions, leading to a biased model that performs perfectly for some hospitals while exhibiting undesirable performance for other sites. The main goal of this section is to improve model "*fairness*" among participating hospitals by proposing a novel federated learning scheme called *Proportionally Fair Federated Learning*, short Prop-FFL. This framework is based on a novel optimization objective function to decrease the performance variations among participating hospitals. This function encourages a fair model, providing us with more uniform performance across participating hospitals. The proposed Prop-FFL is validated on two histopathology datasets as well as two general datasets to shed light on its inherent capabilities. The experimental results suggest promising performance in terms of learning speed, accuracy, and fairness.

4.1 Introduction

The main concern of this section is how to aggregate the training results at the central server. Most existing methods train the global model by simply minimizing the average training losses of all local centers. However, these methods do not provide a performance

guarantee for each individual hospital since they focus on the average training results. These methods lead to a global model that its performance varies among all the participants [26]. This problem is expected to worsen in real-world scenarios where the data from different medical centers exhibit heterogeneity both in terms of size and distribution.

While the main goal is to cope with the data heterogeneity by modifying the aggregation of training results at the central server, another approach that deals with highly non-IID data distributions is *personalized federated learning* (PFL) [54], proposed to improve individual performances by enabling each client to learn a customized model. However, one major drawback of a PFL model is the loss of generalization since the ultimate goal is to learn a separate model for each hospital. While PFL can be helpful for some personalized health applications, it is certainly not suitable for other applications requiring a generalized model to avoid disparity in healthcare delivery [55]. Besides, PFL requires further analysis to adjust model parameters for each hospital, which increases the overhead. Therefore, compared with PFL, modifying the aggregation rule at the central server can cope with non-IID data distribution and improve generalization without extra overhead.

The proportional fair resource allocation of wireless communication systems is adopted to introduce *proportionally fair federated learning*, short Prop-FFL, a novel optimization objective in federated learning, to boost more uniform model performance across participating hospitals. The main goal of Prop-FFL is to train a fair model, which is not biased toward any of the participating hospitals at the expense of having undesirable performance for other local centers. To the best of the author's knowledge, this is the first time that proportional fairness is formulated to modify the deep learning train objective.

The related literature on fairness in federated learning is reviewed in the next section. The intuition behind Prop-FFL is described and formulated in Section 4.3, providing a detailed explanation of the proposed method. In Section 4.4, various experimental results are provided to evaluate Prop-FFL on two histopathology datasets. Also, additional experimental results are provided on two general datasets to verify the effectiveness of Prop-FFL. Finally, the proposed method is concluded in Section 4.5.

4.2 Background: Fairness in Federated Learning

Most existing federated learning methods focus on the average performance of the global model. However, this might lead to a highly variable performance among participant hospitals. For example, relying on these methods make the model biased towards hospitals with larger datasets, having undesirable performance on those hospitals that have access

to only small datasets. Additionally, in real-world scenarios, participants may have non-IID data such that learning hosital distributions become a challenge. In these scenarios, focusing on the average performance will result in a model that performs well for some hospitals while having performance degradation on other hospitals. This means that there is no performance guarantee for individual hospitals.

To address this concern, fairness in federated learning has attracted considerable attention recently. The key point of fairness is that *fair sharing is not equal sharing always* [56]. As a result, one can design FL methods to ascertain that the participant hospitals for which the trained model has unacceptable performance on their data *should contribute more to the global loss function*. In this regard, q-fair federated learning has been proposed [57]. The authors introduced a set of novel optimization problems, namely q-fairness, to apply fairness in federated learning. Having the parameter q in their proposed optimization problem enables the central server to guide the loss function more desirably. The authors also introduce q-fairness FedSGD, or q-FedSGD, to apply fairness in FedSGD. According to the literature review, q-fair federated learning [57] is the benchmark for fairness in federated learning, therefore, the proposed method will be compared with q-FedSGD. There are other works that consider fairness in federated learning. For example authors in [58] introduce a novel form of fairness in federated learning, called min-max optimization. The authors address the fairness problem by employing min-max optimization, focusing on the *weakest participant with maximum loss*. However, changing the focus of the model toward the worse performing participant is not reasonable since it may degrade the total performance. The authors in [59] focus on both fairness and robustness of federated learning. They address these challenges by dynamically choosing a fraction of local centers to participate in training. However, this approach cannot be applied in medical domain because there are limited number of participant hospitals as oppose to wireless network applications where there are thousands of mobile users. Authors in [60] address fairness using different approach. They propose a novel collaborative fair federated Learning framework to enforce participants to converge to different models.

4.3 Proportional Fairness

This section provide a detailed explanation of the proposed proportionally fair federated learning (Prop-FFL). While the purpose of the most existing federated learning methods is to minimize the average training loss over all participant hospitals, this aggregation might lead to a variable performance among hospitals, having a bias toward some participants that contributed more toward the loss function, and hence adjusting the model parameters.

To resolve this issue, *fairness should be considered during the aggregation procedure on the central server.*

Inspired by proportional fair resource allocation methods in the wireless network [61], Prop-FFL is introduced to learn a global model that performs well for all participants, not just some of them. The main goal of Prop-FFL is to enable all hospitals to fairly (not necessarily equally) contribute to the training of the global model. Prop-FFL focuses on the participants with rather poor performance, changing the network parameters to improve their performance. The Prop-FFL ascertains fairness by modifying the optimization problem at the central server. It encourages the model not to have undesirable performance on any of hospitals while minimizing the total training loss. The proposed method is the extension of FedSGD since it is computationally efficient for hospitals and has convergence guarantee [25]. The expansion of the proposed method on FedAvg to create more communication efficient framework will be left for future works.

4.3.1 Problem Formulation

Consider a set of hospitals $\mathcal{M}$ with M hospitals with privately labeled data indexed by $\{1, 2, \ldots, M\} \in \mathcal{M}$. Each hospital $i \in \mathcal{M}$ has only access to its local data that is denoted by $\mathcal{D}^i$. Dataset $\mathcal{D}^i$ consists of n_i training data points denoted by $\mathcal{D}^i = \{\mathbf{a}_l^i, b_l^i\}_{l=1}^{n_i}$, where $\mathbf{a}_l^i$ is training input and b_l^i is its label, e.g., a primary diagnosis. The total number of samples of all hospitals is $n = \sum_{i=1}^{M} n_i$. As in most real-world applications, datasets $\mathcal{D}^i\}_{i=1}^{M}$ may not have identical independent distributions (IID).

Prop-FFL applies fairness by encouraging the model to perform similarly on all M hospitals. In Prop-FFL, the optimization objective function composes of two terms, one to use fairness and one to reduce the training loss. The fairness term aims to adjust the model parameters such that all M hospitals have a similar training loss. To achieve that, the fairness term is defined as an optimization problem that maximizes the multiplication of loss functions of all M hospitals. It will be shown that the optimal solution of the fairness term emerges when all the M hospitals have the same training loss. The following provide an elaboration on these two terms of the optimization objective function starting with the fairness term. To mathematically explain, the relative training loss of the kth hospital is defined as normalized loss and denoted by

$$F_k'(w) = \frac{F_k(w)}{\sum_{j=1}^{M} F_j(w)}, \quad 1 \leq k \leq M, \tag{4.1}$$

where w is the global model parameters and $F_k(w)$ is the training loss function of the kth hospital. Obviously, $0 < F'_k(w) < 1$, and $\sum_{k=1}^{M} F'_k(w) = 1$. Next, $F(w)$ is defined as the multiplication of hospitals' relative losses given by

$$F(w) = \prod_{k=1}^{M} p_k F'_k(w), \tag{4.2}$$

where p_k is the probability associated with the number of samples of the kth hospital and

$$\sum_{k=1}^{M} p_k = 1$$

. p_k specifies the amount of contribution of the kth hospital in loss function. In the experiments, it is assumed that

$$p_k = \frac{n_k}{n}.$$

To establish fairness, it is required to maximize $F(w)$. To make the optimization problem more computationally convenient, one can take the log of both sides of (4.2) to convert the multiplication to summation. Since the log function is strictly increasing, this conversion does not change the optimal solution. Therefore, (4.2) will be rewritten as

$$\begin{aligned} \log\left(F(w)\right) &= \log\left(\prod_{k=1}^{M} p_k F'_k(w)\right) \\ &= \sum_{k=1}^{M} \log\left(p_k F'_k(w)\right), \end{aligned} \tag{4.3}$$

and the optimization problem at the central server will be

$$\begin{aligned} \max_{w} \quad & \sum_{k=1}^{M} \log\big(p_k F'_k(w)\big), \\ \text{s.t.} \quad & F'_k(w) = \frac{F_k(w)}{\sum_{j=1}^{M} F_j(w)}. \end{aligned} \tag{4.4}$$

The maximum operator in (4.4) does not aim to increase the amount of loss functions. Since $\sum_{k=1}^{M} F'_k(w) = 1$, the maximization problem cannot violate this constraint

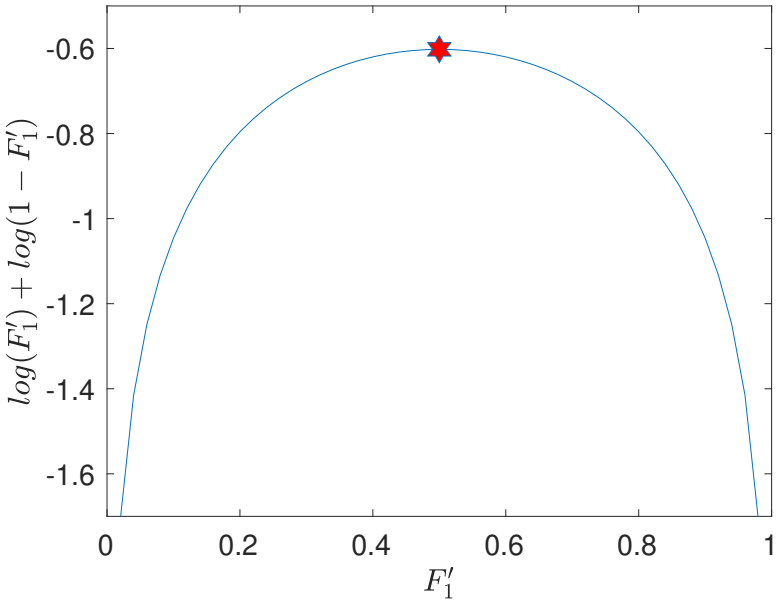


Figure 4.1: The function plot of (4.3) when $M = 2$ and $p_1 = p_2 = 0.5$. The max of $log(F_1') + log(1 - F_1')$ happens in $F_1' = 0.5$. This means that the maximum occurs when both hospitals achieve the same relative loss, $F_1' = F_2' = 0.5$.

and increase the loss functions. Instead, the maximum of $F(w)$ occurs when all hospitals approach the same relative loss while the summation of the relative losses is 1. In other words, the optimal solution of (4.4) happens when $F_k'(w) = \frac{1}{M}$. The proof for that is provided in Appendix A. This means all M hospitals achieve the same training loss, $F_1(w) = F_2(w) = \cdots = F_M(w)$, which is the fairness ultimate goal of Prop-FFL.

The proportional fairness optimization problem (4.4) could be more clear by two examples. Figure. 4.1, illustrates the function plot of (4.3) when $M = 2$, $\log(F_1'(w)) + \log(F_2'(w))$ where $F_2'(w) + F_1'(w) = 1$. As can be seen, the maximum occurs when both participants get the same portion of unit 1, meaning $F_1'(w) = F_2'(w) = 0.5$. As another example, consider 4 hospitals that collaborate in the federated learning fashion. Using the proposed proportional fairness, each of those four hospitals would fairly contribute to learning the global model, achieving the same relative loss $F_k'(w) = 0.25, \quad 1 \leq k \leq 4$.

The optimization problem in (4.4) can be simplified as

$$\begin{aligned}\max_w \sum_{k=1}^{M} \log\left(p_k F_k'(w)\right) &= \min_w \sum_{k=1}^{M} -\log(p_k F_k'(w)) \\ &= \min_w \sum_{k=1}^{M} -\log\left(p_k \frac{F_k(w)}{\sum_{j=1}^{M} F_j(w)}\right) \\ &= \min_w \sum_{k=1}^{M} \log\left(\frac{1}{p_k}\frac{\sum_{j=1}^{M} F_j(w)}{F_k(w)}\right) \\ &= \min_w c + \sum_{k=1}^{M} \log\left(\frac{\sum_{j=1}^{M} F_j(w)}{F_k(w)}\right),\end{aligned} \tag{4.5}$$

where $c = \sum_{k=1}^{M} \log\left(\frac{1}{p_k}\right)$ is a constant. Since c is not a function of w, one can omit that from the objective function. Therefore, the objective function is

$$\min_w \mathcal{L}(w) = \sum_{k=1}^{M} \underbrace{\log\left(\frac{\sum_{j=1}^{M} F_j(w)}{F_k(w)}\right)}_{G_k(w)}. \tag{4.6}$$

For notation simplicity, $G_k(w)$ is defined as

$$G_k(w) = \log\left(\frac{\sum_{j=1}^{M} F_j(w)}{F_k(w)}\right).$$

The optimization problem in (4.6), intends to apply fairness and adjust the model's parameter such that all the hospitals get the same relative loss. However, this objective function does not decrease the training loss. To reduce the training loss, the objective function $\mathcal{Q}(w)$ is adopted from [57]:

$$\min_w \mathcal{Q}(w) = \sum_{k=1}^{M} \frac{1}{q+1} F_k^{q+1}(w), \tag{4.7}$$

where q is a constant that will be tuned in the experiments, and $F_k(w)$ is the training loss function of the kth hospital. The final objective function that reduces the total training

loss while imposes fairness is a convex combination of objective functions (4.7) and (4.6) as follows:

$$w^* = \arg\min_w (1-\lambda) \underbrace{\sum_{k=1}^{M} \frac{1}{q+1} F_k^{q+1}(w)}_{\text{term I:}\mathcal{Q}(w)} + \lambda \underbrace{\sum_{k=1}^{M} \log\left(\frac{\sum_{j=1}^{M} F_j(w)}{F_k(w)}\right)}_{\text{term II:}\mathcal{L}(w)}. \tag{4.8}$$

The first term in (4.8), term I, aims to reduce the total loss. The second term in (4.8), term II, applies proportional fairness among participants, forcing relative losses to be close to each other. The hyperparameter $0 < \lambda < 1$ adjusts the level of fairness in the objective function.

Since term I and term II in (4.8) are differentiable and smooth, the optimization problem (4.8) can be solved by adopting the stochastic gradient decent (SGD). To apply SGD, it is needed to compute the derivative of the objective function with respect to w. The derivative of $\mathcal{Q}(w)$ with respect to w is

$$\nabla \mathcal{Q}(w) = \sum_{k=1}^{M} F_k^q(w) \nabla F_k(w), \tag{4.9}$$

and derivative of $\mathcal{L}(w)$ with respect of w can be given as

$$\nabla \mathcal{L}(w) = \sum_{k=1}^{M} \nabla G_k(w), \tag{4.10}$$

where $\nabla G_k(w)$ is

$$\nabla G_k(w) = \nabla \log \left(\frac{\sum_{j=1}^{M} F_j(w)}{F_k(w)} \right) = \frac{\sum_{j=1}^{M} \big(F_k(w) \nabla F_j(w) - \nabla F_k(w) F_j(w) \big)}{\sum_{j=1}^{M} F_k(w) F_j(w)}. \tag{4.11}$$

Finally, given the gradients in (4.9), (4.10), and (4.11), the gradient of the objective function (4.8) can be formulated as

$$\Delta = \sum_{k=1}^{M} (1-\lambda) F_k^q(w) \nabla F_k(w) + \lambda \nabla G_k(w). \tag{4.12}$$

The gradient in (4.12) allows us to aggregate the updates from all hospitals in a proportional fair setting using SGD, estimating the directions to update the model parameters.

4.3.2 Proportionally Fair Federated Learning: Prop-FFL

Similar to FedSGD, in Prop-FFL, participating hospitals train the global model using one batch of their local training data. Next, each hospital feeds back the loss function and the gradient of the loss function to the central server. Finally, relying on Prop-FFL, the central server aggregates the training results, updating the model parameters. These steps are repeated until convergence. More details of Prop-FFL have been provided in Algorithm 5.

4.4 Evaluation

This section presents the experimental results of the proposed Prop-FFL. The effectiveness of Prop-FFL has been verified on four publicly available datasets.

Algorithm 5 Proposed Proportionally fair federated learning

Input: M, T, w^0, η, λ
Output: w^{T-1}

1: **for** $t = 0, \ldots, T-1$ **do**
2: Server sends w^t to all hospitals
3: Each hospital k computes $F_k(w^t)$ and $\nabla F_k(w^t)$ for one batch of data
4: Hospitals send $F_k(w^t)$ & $\nabla F_k(w^t)$ back to server
5: Server computes $\nabla G_k(w^t)$ based on (4.11)
6: Server updates w^{t+1} as

$$\Delta_k^t = (1-\lambda)F_k^q(w^t)\nabla F_k(w^t) + \lambda \nabla G_k(w^t)$$

$$w^{t+1} = w^t - \eta \sum_{k=1}^{M} \Delta_k^t$$

7: **end for**
8: **return** w^{T-1}

Table 4.1: Summary of datasets

Dataset	**# Participants**	**# Samples**
MNIST (non-medical)	10	60,000
FMNIST (non-medical)	10	60,000
Histopathology-Kidney	4	642,277
Histopathology-Lung	6	303,033

4.4.1 Image Datasets

This section provides detailed explanations on the datasets that have been used in the experiments. Prop-FFL will be studied on two histopathology datasets. Also, additional experimental results will be provided on two non-medical datasets to confirm the effectiveness of the proposed method for well-known datasets. The statistical summary of all four datasets has been provided in Table 4.1. In all datasets, data has been divided into 60%, 20%, 20% groups, for training, validation, and testing, respectively. Additional details of datasets and models used in the experiments are described below.

MNIST

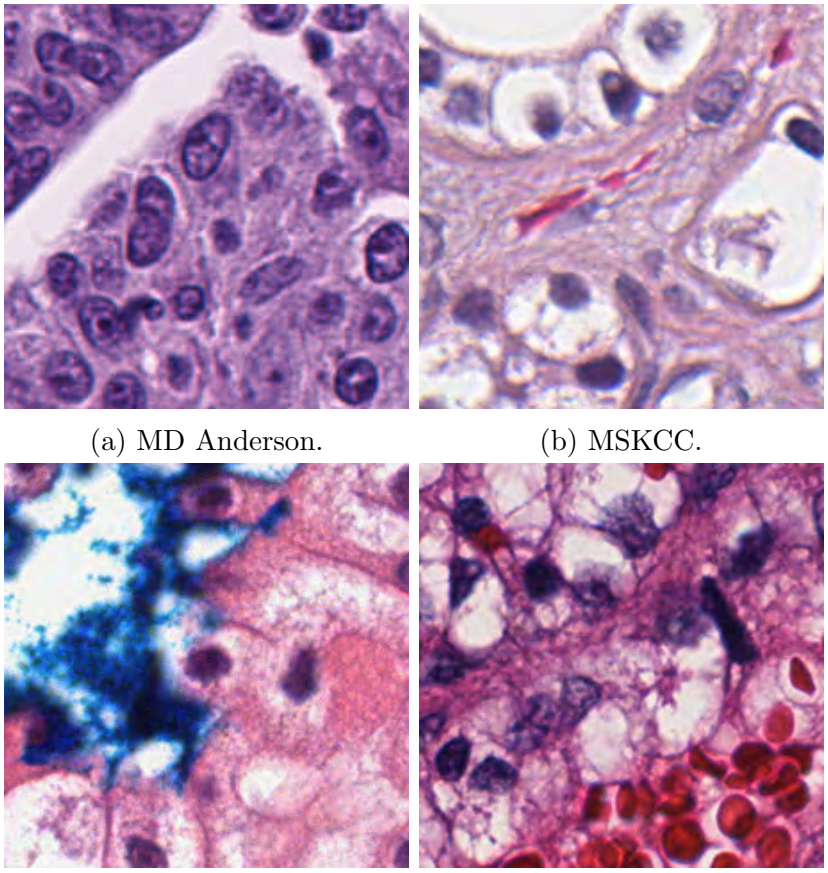
This dataset is a well-known and widely collection of hand-written digits. The images are black and white of size 28×28 pixels [62]. To investigate the behaviour of the proposed federated learning method, it is assumed 10 participants in MNIST with non-IID data distribution. The approach in [25] is adopted to distribute data samples between participants in a non-IID fashion. First, data samples are sorted based on the digit labels, dividing it into 20 shards of size 3,000 samples each, and then randomly assign 2 shards to each of 10 participants. This partitioning provides us with non-IID data distribution as participants mostly have two digits. For the classification model, a convolutional neural network (CNN) is considered with two 5×5 convolution layers, each followed with 2×2 max pooling and ReLu activation function. Then, two fully connected layers were used with the first one being followed by the ReLu activation and the second one by a softmax output layer.

FMNIST

This dataset comprises of size 28×28 images from fashion products. They are gray scale and from 10 fashion categories [63]. To study the federated learning on FMNIST, 10 participants are considered. The data distribution approach is exactly the same as what has been established for MNIST previously. The classifier is also the same as what have been used for MNIST.

Histopathology Datasets

For final experiments, The Cancer Genome Atlas (TCGA) [39, 50], the largest public archive of histopathology whole-slide images (WSIs), will be used. TCGA provides researchers with more than $30,000$ H&E stained histopathology WSIs prepared by various medical centers. In TCGA, the variation between hospitals' data occurs in many different aspects. Figure 4.2 and 4.3 represent sample patches from four different hospitals of kidney dataset [64] and six different hospitals of lung histopathology dataset [37] respectively. As can be seen, staining can differ significantly among hospitals depending on staining protocols. Additionally, images from different hospitals can have various artifacts that might be characteristic for each hospital. For example, as Figure 4.2c shows there is a blue ink in the image. Other hospitals might have other artifacts, such as blur, tissue-fold, tears depending on their tissue preparations, imaging protocols, and scanners. While common artifacts are unavoidable during histopathology slide preparation, It is assumed that each

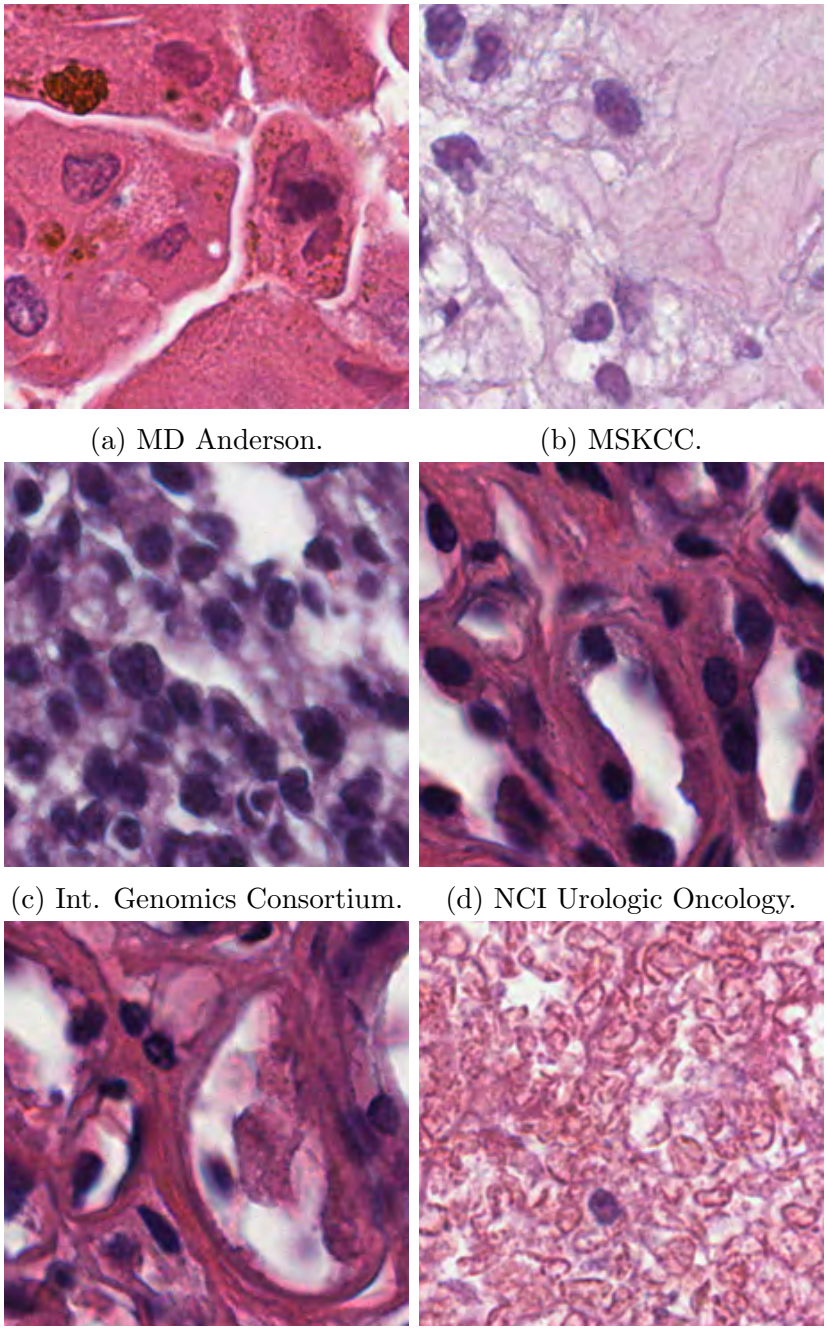


(a) MD Anderson. (b) MSKCC.

(c) Int. Genomics Consortium. (d) NCI Urologic Oncology.

Figure 4.2: Histopathology patch samples from each of the four hospitals in *kidney* datasets.

hospital has a quality control unit and the images have an acceptable quality level that machine learning models can access. This assumption is not specific to the proposed method. It is a very general assumption in machine learning that data is reliable with acceptable quality for learning tasks. Only a few studies have looked at the behavior of a deep model in presence of artifacts. However, the laboratory practice at different sites may make the quality control a more urgent need [65, 66, 67] Those artifacts, as well as different tissue preparation protocols, stain variation, label distribution variability across sites, and using different scanners and protocols may cause non-IID data distribution among hospitals. As recent investigations have reported 'normalization and augmentation do not prevent models from learning site-specific characteristics', which cause non-IID challenges, although stain normalization may help increase the accuracy [68].Our proposed solution copes with



(a) MD Anderson. (b) MSKCC.

(c) Int. Genomics Consortium. (d) NCI Urologic Oncology.

(e) NCI Urologic Oncology. (f) NCI Urologic Oncology.

Figure 4.3: Histopathology patch samples from each of the four hospitals in *lung* datasets.

Table 4.2: The summary of *kidney* histopathology dataset [64]

Client	# Slides	# Patches
C1: MD Anderson Cancer Center	95	138,986
C2: International Genomics Consortium	87	60,149
C3: MSKCC	198	245,041
C4: NCI Urologic Oncology Branch	44	198,101

non-IID data distribution of participant hospitals in a federated learning scenario. The performance of Prop-FFL will be validated for the classification of histopathology images of different hospitals using two histopathology datasets, kidney and lung histopathology datasets. Both lung and kidney datasets are publicly available in TCGA resposiytory with assigned primary diagnoses as labels for the entire image [39].

- *Kidney Datase* [64]. The WSIs diagnosed with *kidney cancer* have been selected from TCGA to construct a dataset of several institutions. The kidney cancer includes three most frequent subtypes, namely
  - Clear cell renal cell carcinomas (ccRCC),
  - Papillary renal cell carcinomas (pRCC),
  - Chromophobe renal cell carcinomas (crRCC).

  Only diagnostic WSIs scanned at a magnification of 40x have been considered from TCGA. This study included hospitals that had a sufficient number of WSIs spanning across all three subtypes (ccRCC, pRCC, crRCC). Only four hospitals met this requirement in TCGA, namely NCI Urologic Center, International Genomics Center, MSKCC, and MD Anderson. Since WSIs are extremely large with high magnification, they have been divided into small patches for further analysis [51]. Readers are referred to [64] for more details on how patches have been extracted from WSIs in this dataset. Table. 4.2 represents the number of WSIs and patches of each hospital in kidney histopathology dataset and Figure.4.5 shows the distribution of kidney cancer sub-types in each hospital. For the kidney histopathology image classification model, first pretrained DenseNet121 [52] is employed to extract image features of length 1,024. Next, a fully connected layer followed by ReLu activation function is used. Then, three fully connected layers are followed by the softmax output layer.

- *Lung Dataset* [37]. This dataset includes TCGA WSIs diagnosed with non-small cell lung cancer (NSCLC) which has two frequent subtypes, namely

Table 4.3: The summary of *lung* histopathology dataset [37].

Client	**# Slides**	**# Patches**
C1: International Genomics Consortium	267	66,483
C2: Indivumed	211	52,539
C3: Asterand	207	51,543
C4: Johns Hopkins	199	49,551
C5: Christiana Healthcare	223	55,527
C6: Roswell Park	110	27,390

– Lung Adenocarcinoma (LUAD)

– Lung Squamous Cell Carcinoma (LUSC).

In TCGA, there are only six hospitals that have a sufficient number of WSIs from both LUAD and LUSC subtypes. The statistics of this dataset for each hospital are presented in Table. 4.3 and Figure. 4.4. WSIs have been divided into patches of size 1000×1000 pixels. More details on patch extraction and selection are provided in [37]. For the classification of lung histopathology WSIs, first pretrained DenseNet121 [52] is employed to extract image features of length 1,024. Next, attention gated multiple instance learning (MIL) is used to effectively combine the patch feature vectors of each WSI and create a feature of size 1024 for each WSI [53]. More details about this MIL network is provided in [53]. Next, a fully connected layer is used followed by ReLu activation function. Finally, three fully connected layers are applied followed by softmax output layer.

Patch-level diagnosis is used for the kidney dataset and slide-level diagnosis for the lung dataset. There are two reasons why patch-level for one and slide-level for the other was used. First, datasets introduced in the literature and a similar classifier for each were employed. For example, many authors use a MIL classifier for the kidney dataset that trains the network based on the slide-level annotation. However, some authors have used another classifier for the lung dataset, which trains the classifier on patch bases. The second reason is that the number of slides in the fourth hospital in the kidney dataset is insufficient to perform the slide-level evaluation. However, there are enough lung slides in each hospital to report slide-level performance.

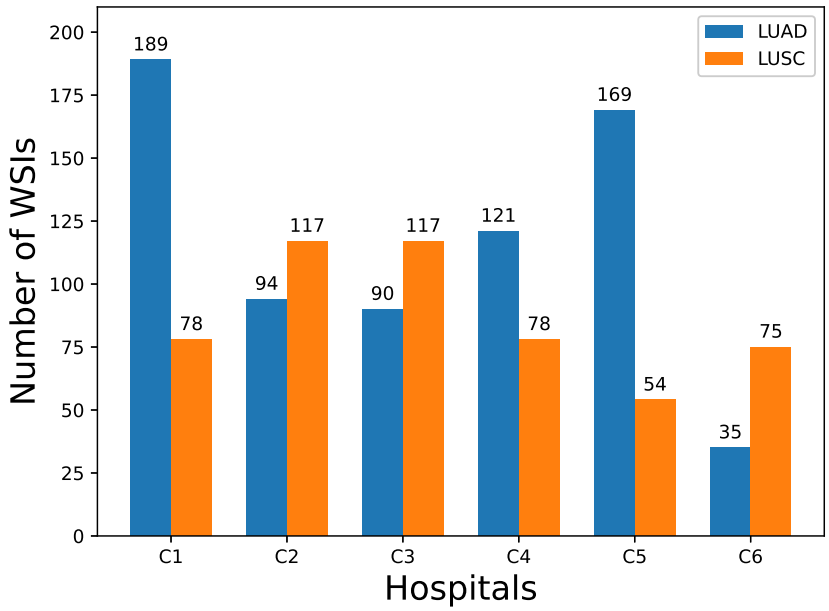


Figure 4.4: The distribution of the classes in *lung* dataset.

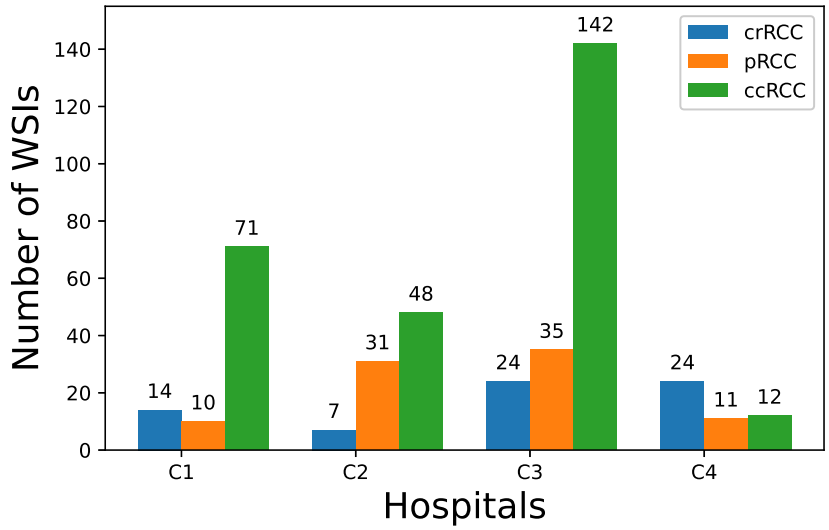


Figure 4.5: The distribution of the classes in *kidney* dataset.

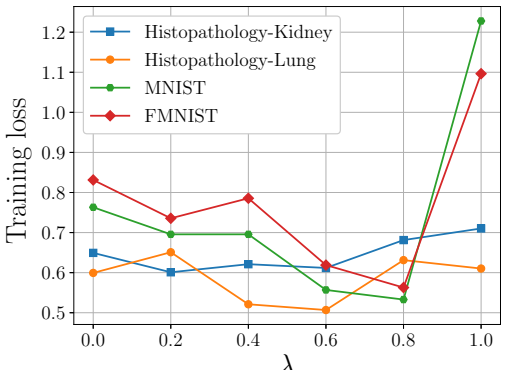
Table 4.4: The accuracy of each hospital in each method for kidney dataset.

Hospitals	FedSGD	q-FedSGD	Prop-FFL
Hos1	71.23 ± 0.12	76.94 ± 0.56	78.84 ± 0.83
Hos2	72.01 ± 0.04	75.78 ± 0.63	77.12 ± 0.58
Hos3	66.87 ± 0.11	75.94 ± 0.71	76.21 ± 0.56
Hos4	82.41 ± 0.08	83.89 ± 0.45	84.36 ± 0.35
Var	32.54	11.22	10.00

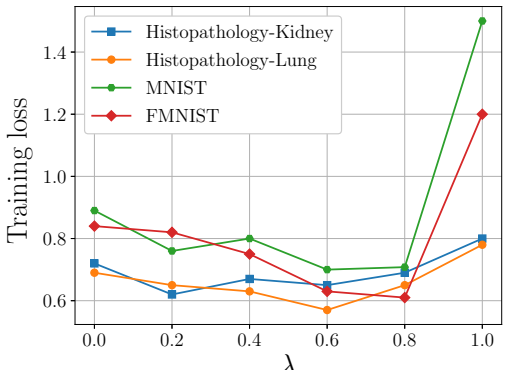
4.4.2 Impact of λ

The final loss function is a convex combination of two loss functions in (4.8), one to reduce the loss and the other one to apply fairness. The parameter $0 < \lambda < 1$ in (4.8) determines the weight of each of those two loss functions. It is needed to investigate the impact of λ on the performance. Figure 4.6 depicts the training loss and testing accuracy vs. λ using training and validation datasets, respectively. The results are obtained after 100 epochs of training. The learning rate and q in (4.8) have been tuned for each dataset to gain the highest accuracy on the validation dataset. This study provides us with insight into the behaviour of Prop-FFL as well as with the default value of λ in order to enable full automation. According to Figure 4.6, $\lambda = 0.6$ has an acceptable performance on all datasets. Therefore, $\lambda = 0.6$ is considered as the default value for all experiments in the next section. $\lambda = 0.6$ means that more focus is on part of the loss function that applies fairness. $\lambda = 0.6$ means that more focus is on the part of the loss function that applies fairness. However, since the nature of the data, data distribution, the number of clients, etc are different between datasets, to gain the best possible performance this parameter were also specifically tuned for each dataset, which will result in better performance compared to the default value.

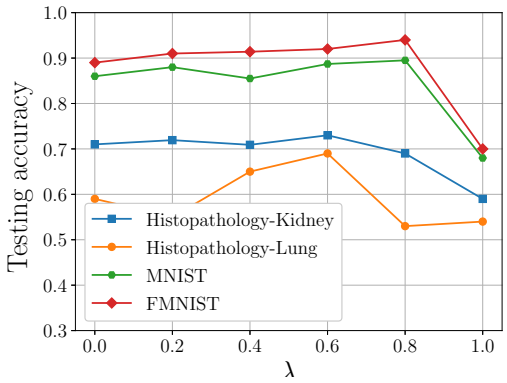
According to Figure 4.6, as expected, large λ values ($\lambda \to 1$) degrades the performance. The reason is that $\lambda \to 1$ diminishes term I in (4.8), leaving us only with term II as the objective function. Since term II in (4.8) applies only proportional fairness and cannot guarantee the decrease of the training loss, the results for $\lambda \to 1$ are rather undesirable. For small λ values ($\lambda \to 0$) the performance is better than large λ values. The reason is that $\lambda \to 0$ reduces the objective function in (4.8) to term I which decreases training loss. Therefore, since it guarantees the training loss reduction, its performance is more acceptable compared to $\lambda \to 1$.



(a) Average training loss vs. λ.



(b) Training loss of worst hospital.



(c) Average accuracy vs. λ

Figure 4.6: Impact of λ on training loss and accuracy.

Table 4.5: The accuracy of each hospital in each method for lung dataset.

Hospitals	FedSGD	q-FedSGD	Prop-FFL
Hos1	66.22 ± 0.02	66.68 ± 1.49	72.23 ± 1.27
Hos2	58.06 ± 0.01	73.16 ± 1.52	76.81 ± 0.59
Hos3	50.14 ± 0.01	62.74 ± 1.57	74.57 ± 1.41
Hos4	74.19 ± 0.01	72.64 ± 1.95	74.50 ± 0.93
Hos5	71.20 ± 0.02	70.96 ± 1.62	76.44 ± 1.28
Hos6	40.13 ± 0.001	52.10 ± 2.08	60.81 ± 1.82
Var	143.54	53.89	29.84

4.4.3 Experiments and Results

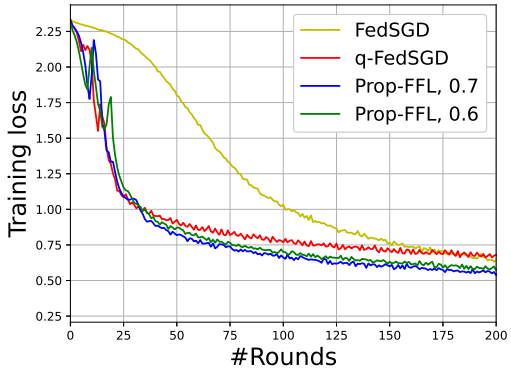
Without loss of generality, it is assumed that all hospitals/users can participate in learning the global model without having any delay and network issue. According to the study in Section 4.4.2, $\lambda = 0.6$ provides us with acceptable performance on all four datasets (medical and non-medical). Therefore, all results for Prop-FFL have been reported for $\lambda = 0.6$ as well as for the best performing λ in each case. The latter requires additional computational overhead to identify the best lambda value for each dataset. Prop-FFL will be compared with two other benchmark methods, FedSGD and q-FedSGD [57]. Learning rate and parameter q have also been fine tuned to get the best performance for each method. In the experiments, 10^{-10} was added to the input of the *log* function to prevent occurrence of undefined value in logarithmic function.

As shown in Figure 4.7 and 4.9, the training has converged for the general datasets, MNIST and FMNIST, as the training loss and testing accuracy have reached stable behavior. Proposed method could gain better final accuracy as well as a faster convergence rate compared to other methods. For histopathology datasets, proposed method exhibits faster convergence, although due to time considerations, the training was stopped after 300 rounds of weight adjustments. In Figure 4.7 and 4.8, the performance of Prop-FFL is evaluated on MNIST dataset. These figures demonstrate the performance in terms of training loss and testing accuracy. The batch of size 1024 is used. $\lambda = 0.7$ was the best value for MNIST dataset. Therefore, the results are represented for both $\lambda = 0.7$ and the default value $\lambda = 0.6$. Figure 4.7a and 4.7b represent the average training loss and the worst training loss of ten participants, respectively. Both q-FedSGD and Prop-FFL outperform FedSGD in these two figures. Prop-FFL performs better than q-FedSGD specially in Figure 4.7b. The standard deviation of training loss of all ten participants has been shown in Figure 4.7c. As can be seen, Prop-FFL and q-FedSGD perform close and better than FedSGD. Figure 4.8a and 4.8b depicts the average testing accuracy and the worst testing accuracy of all ten participants, respectively. As can be seen, Prop-FFL outperforms q-FedSGD. They both surpass FedSGD. Figure 4.8c represents the standard deviation of testing accuracy of all users. As can be seen, Prop-FFL is slightly better than q-FedSGD and both outperform FedSGD. Figure 4.9 and 4.10 represent the performance of Prop-FFL on FMNIST dataset. All experimental considerations are similar to what have been done for MNIST. For FMNIST dataset, Prop-FFL has its best performance for $\lambda = 0.8$. Therefore, the results have been depicted for both $\lambda = 0.6, 0.8$.

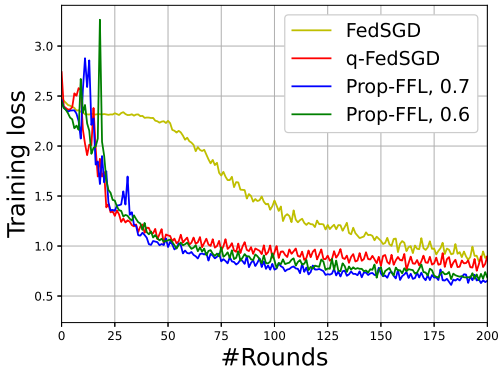
Figure 4.11, 4.12 depict the performance evaluation on histopathology-kidney dataset. The best performing λ on this dataset is $\lambda = 0.1$. Therefore, the results have been provided for $\lambda = 0.1, 0.6$. The batch of size 1024 is used for this dataset. Figure 4.11a represents the

average training loss of all four hospitals, and Figure 4.12a represents the average testing accuracy of all four hospitals. As can be seen, Prop-FFL and q-FedSGD outperform FedSGD. The reason is that both Prop-FFL and q-FedSGD have modified the objective function, providing them with a more generalized global model that is not biased toward any hospital. Figure 4.11b and 4.12b represent the training loss and testing accuracy of the worst performing hospital, respectively. As illustrated, Prop-FFL outperforms q-FedSGD, and q-FedSGD performance exceeds FedSGD. The reason is that the objective function does not allow poor performance for any hospitals. Figure 4.11c and 4.12c depict the standard deviation of training loss and testing accuracy of all hospitals, respectively. As illustrated, the standard deviation of Prop-FFL is smaller than q-FedSGD, and both smaller than FedSGD. The reason is that q-FedSGD and Prop-FFL have modified the objective function, encouraging a fair model to decrease the performance variation across hospitals. Figure 4.13, 4.14 shows the performance on histopathology-lung dataset. The best performing λ on this dataset is same as the default value for λ. Therefore, the results have been depicted for only $\lambda = 0.6$. The batch size of 72 is used for this dataset. As shown, the performance of Prop-FFL is better than q-FedSGD and FedSGD by large margin in this dataset compared to histopathology-kidney dataset. Also, FedSGD have poor performance on this dataset. One possible reason for that might be the number of hospitals which is larger in lung dataset. Table 4.4 and 4.5 represent the test accuracy per hospital for kidney and lung datasets, respectively. The results have been provided by taking an average over five independent experiments. As can be seen, Prop-FFL provides us with more uniform testing accuracy over participant hospitals as the variance of the testing accuracy of hospitals in Prop-FFL is less than the other two methods.

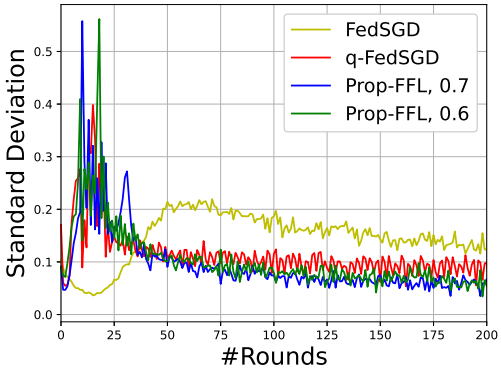
The communication overhead of Prop-FFL and q-FedSGD is the same as FedSGD since both have been built on top of the FedSGD. The only difference between these three methods is the aggregation rule at the central server, which does not impact the communication overhead. Prop-FFL was applied on FedAvg. The results are presented in Appendix B. As can be seen, the concept of fairness does not change the FedAvg performance. This result was expected as averaging moves toward the same results for all hospitals, whereas fairness does emphasize individual contributions. Although FedAvg can improve communication overhead as hospitals do not need to communicate with the central server for each batch of data, FedSGD approach has moderate communication costs in some situations where participant hospitals have small datasets. The reason is that the communication cost of FedSGD is proportional to the number of batches of local datasets. As the total number of batches is small when local datasets are small, FedSGD is more communication efficient. Having small datasets is quite common in medical domain. Therefore, Prop-FFL can provide us with fairness at tolerable communication costs.



(a) $\overline{\text{TL}}$ of all users

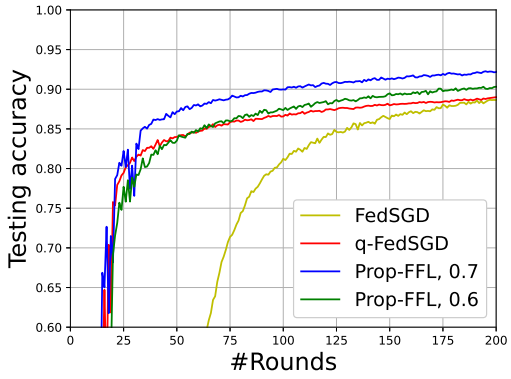


(b) TL of the worst user

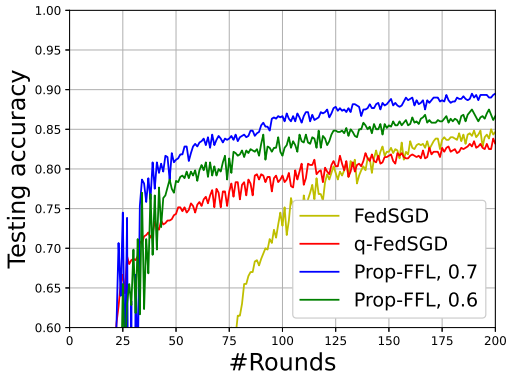


(c) standard deviation of all TLs

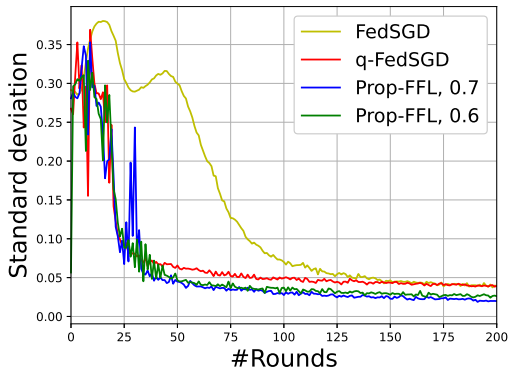
Figure 4.7: Evaluation on **MNIST** dataset with Non IID data distribution. The results for Prop-FFL have been provided for $\lambda = 0.6$ and the best λ. (TL= training loss, $\overline{\text{TL}}$=average training loss), TA= testing accuracy, $\overline{\text{TA}}$=average testing accuracy)



(a) $\overline{\text{TA}}$ of all users

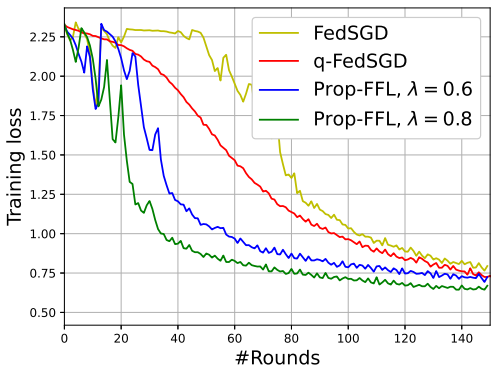


(b) TA of the worst user.

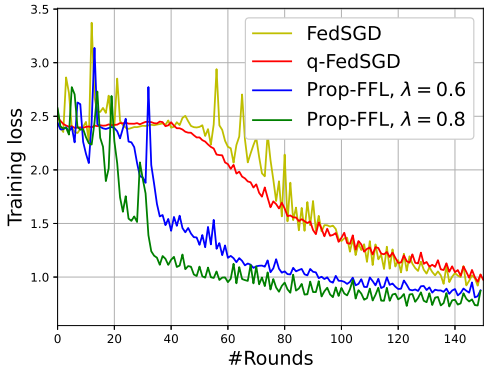


(c) standard deviation of all TAs

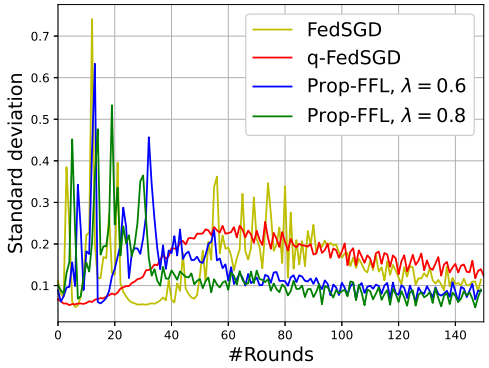
Figure 4.8: Evaluation on **MNIST** dataset with Non IID data distribution. The results for Prop-FFL have been provided for $\lambda = 0.6$ and the best λ. (TL= training loss, $\overline{\text{TL}}$=average training loss), TA= testing accuracy, $\overline{\text{TA}}$=average testing accuracy)



(a) $\overline{\text{TL}}$ over all users

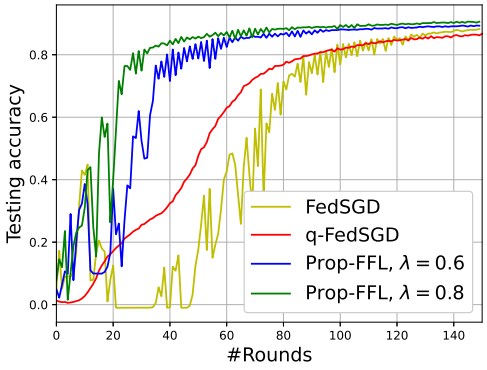


(b) TL of the worst user

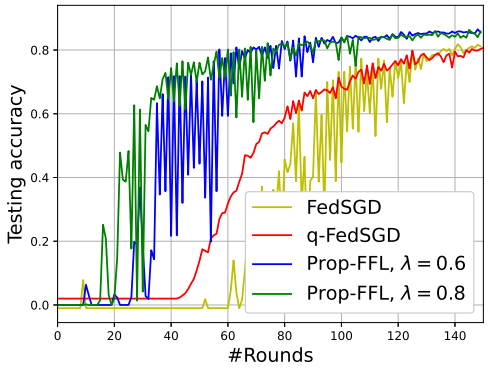


(c) standard deviation of all TLs

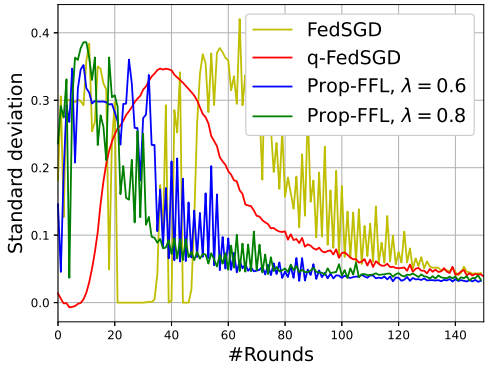
Figure 4.9: Evaluation on **FMNIST** dataset with Non IID data distribution. The results for Prop-FFL have been provided for $\lambda = 0.6$ and the best λ. (TL= training loss, $\overline{\text{TL}}$=average training loss), TA= testing accuracy, $\overline{\text{TA}}$=average testing accuracy)



(a) $\overline{\text{TA}}$ of all users

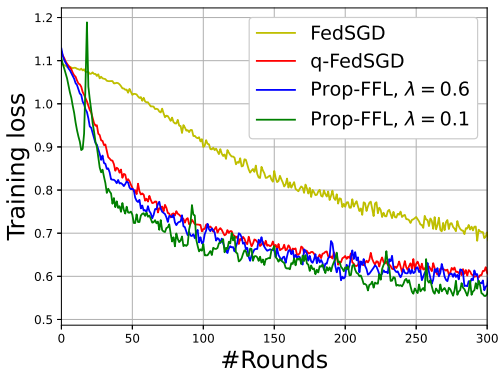


(b) TA of the worst user

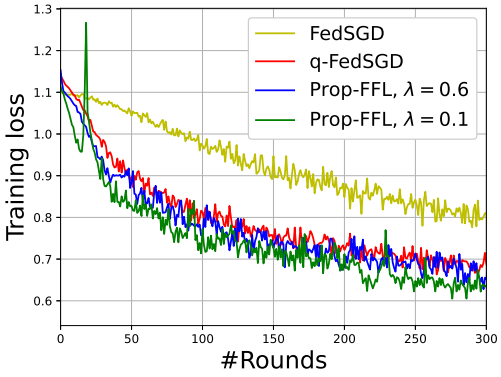


(c) standard deviation of all TAs.

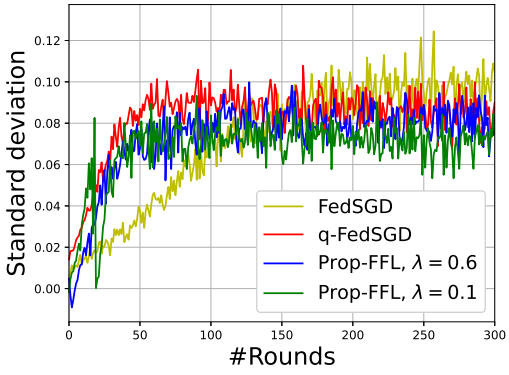
Figure 4.10: Evaluation on **FMNIST** dataset with Non IID data distribution. The results for Prop-FFL have been provided for $\lambda = 0.6$ and the best λ. (TL= training loss, $\overline{\text{TL}}$=average training loss), TA= testing accuracy, $\overline{\text{TA}}$=average testing accuracy)



(a) $\overline{\text{TL}}$ over all hospitals

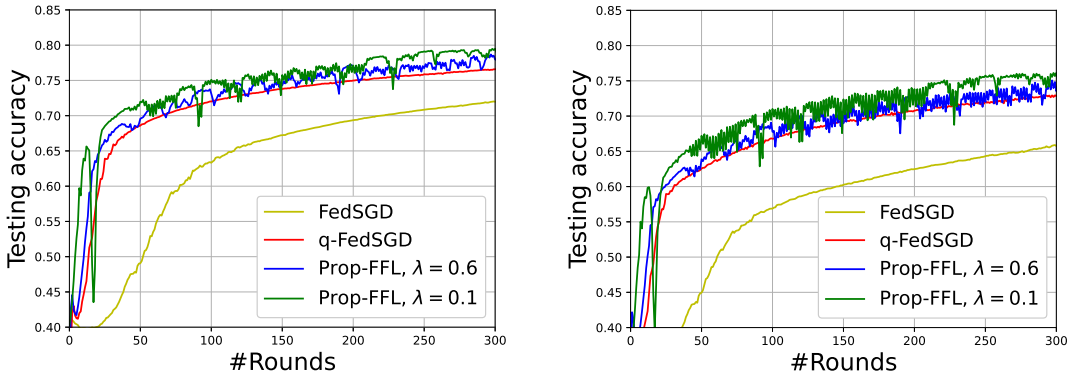


(b) TL of the worst hospital



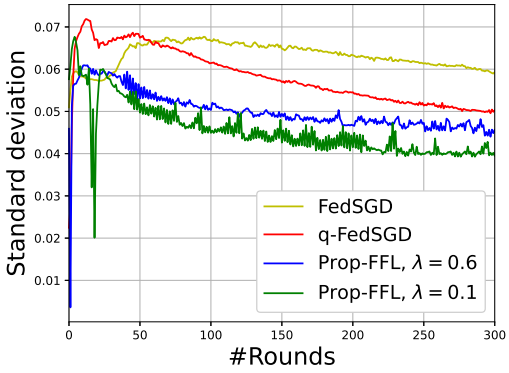
(c) TL variance of all hospitals

Figure 4.11: Evaluation on **Histopathology-Kidney** dataset. The results for Prop-FFL have been provided for $\lambda = 0.6$ and the best λ. (TL= training loss, $\overline{\text{TL}}$=average training loss), TA= testing accuracy, $\overline{\text{TA}}$=average testing accuracy)



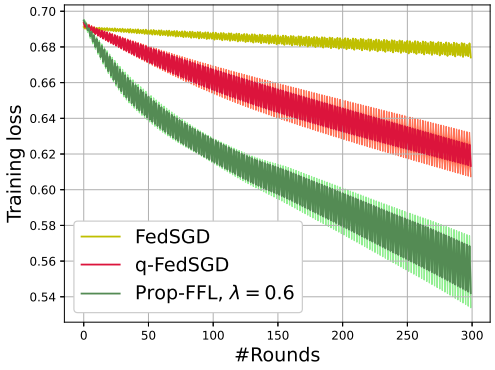
(a) $\overline{\text{TA}}$ of all hospitals

(b) TA of the worst hospital

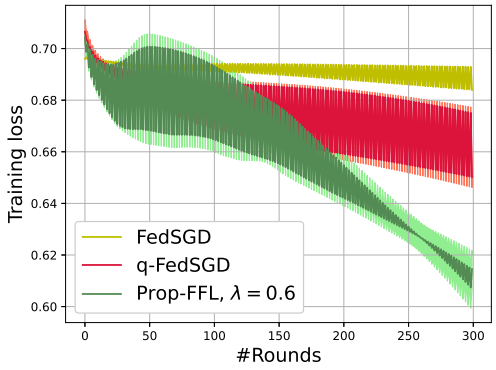


(c) TA variance of all hospitals

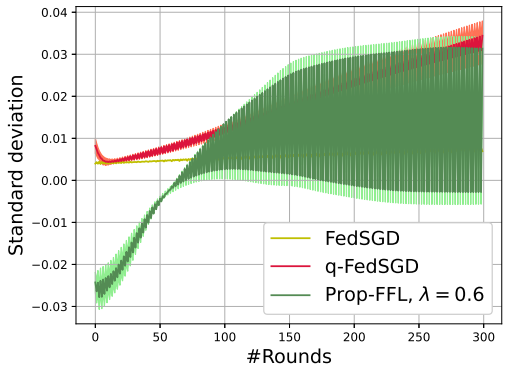
Figure 4.12: Evaluation on **Histopathology-Kidney** dataset. The results for Prop-FFL have been provided for $\lambda = 0.6$ and the best λ. (TL= training loss, $\overline{\text{TL}}$=average training loss), TA= testing accuracy, $\overline{\text{TA}}$=average testing accuracy)



(a) $\overline{\text{TL}}$ over all hospitals

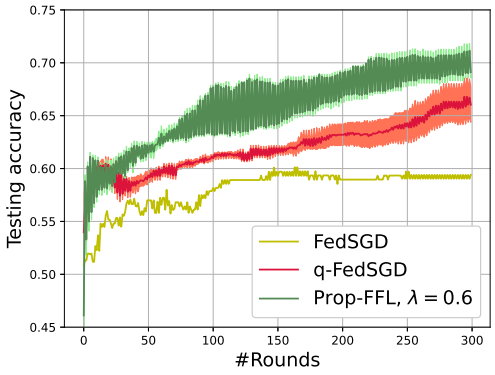


(b) TL of the worst hospital

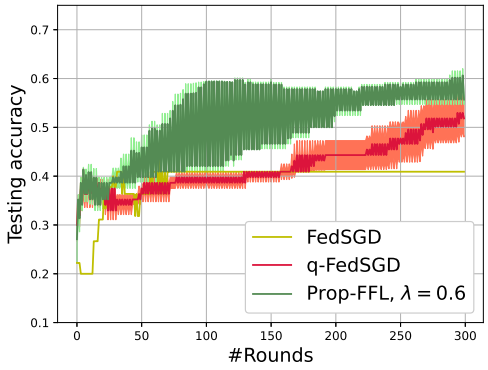


(c) TL variance of all hospitals

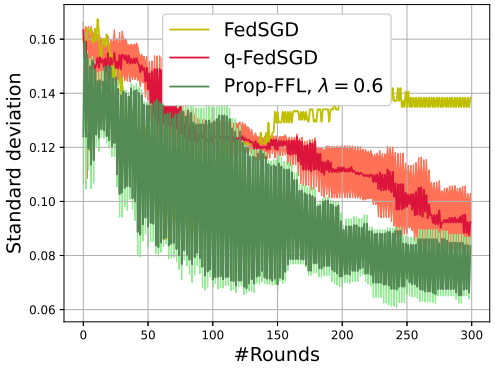
Figure 4.13: Evaluation on **Histopathology-Lung** dataset. The results for Prop-FFL have been provided for $\lambda = 0.6$ which is the best λ. (TL= training loss, $\overline{\text{TL}}$=average training loss), TA= testing accuracy, $\overline{\text{TA}}$=average testing accuracy)



(a) $\overline{\text{TA}}$ of all hospitals



(b) TA of the worst hospital



(c) TA variance of all hospitals

Figure 4.14: Evaluation on **Histopathology-Lung** dataset. The results for Prop-FFL have been provided for $\lambda = 0.6$ which is the best λ. (TL= training loss, $\overline{\text{TL}}$=average training loss), TA= testing accuracy, $\overline{\text{TA}}$=average testing accuracy)

4.5 Summary

In this chapter, the fairness aspect of federated learning among participants was addressed. Proportionally fair federated learning, or Prop-FFL, was proposed to allow all hospitals to fairly contribute to the training of a global model. The intuition behind the proposed idea was explained and mathematical formulations of Prop-FFL were provided. The effectiveness of Prop-FFL has been demonstrated on two histopathology datasets as well as two non-medical datasets to gain insight into its behavior. The experimental results revealed that the proposed method outperforms other federated approaches. In Prop-FFL, similar to FedSGD, hospitals do not need to update model parameters. In future works, one may modify Prop-FFL such that hospitals would be able to update the model parameters, similar to FedAvg, creating more communication efficient framework. Additionally, the fixed step size η was used to update model weights. Employing the second-order SGD to estimate the best step size in each iteration would perhaps improve the convergence.

Chapter 5

Summary and Conclusions

The adoption of digital pathology is expected to revolutionize healthcare systems over the next few years thanks to the advent of digital scanners and the availability of whole slide images (WSIs). Digital pathology may not only help medical centers to improve routine workload and relieve pathologists of laborious and time-consuming tasks but also can enable the research community to apply artificial intelligence to complex diagnostic, prognostic and predictive tasks in medicine. Sophisticated machine learning techniques have the potential to create a second opinion system that facilitates consensuses building to improve diagnosis.

The key enabler of machine learning techniques is large and diverse medical data. However, collecting a massive amount of patients' records from various medical centers in one central location for machine training purposes is prohibitive due to privacy concerns and regulatory restrictions. To address this challenge, federated learning, an emerging technology, may enable a novel collaborative distributed learning paradigm in healthcare. It has opened a new horizons in machine learning for data-sensitive domains, such as healthcare.

Federated learning allows hospitals to keep their data local and perform model training without touching onsite data. Under this framework, hospitals perform training on their local data and share the training results with other hospitals or the central server to collaboratively learn a single machine learning model. In federated learning, data is always local never leaving hospitals.

Federated learning adoption in real-world scenarios, however, exhibits many challenges that need to be addressed. This thesis focused on two challenges of federated learning in healthcare. Two frameworks are proposed to enable hospitals, and the central server

to communicate securely. Additionally, a novel optimization objective problem was introduced to address the non-IID challenge across hospitals. The proposed method modifies the aggregation rule at the central server, improving fairness. The effectiveness of the proposed methods was validated by conducting experiments on WSIs obtained from public TCGA dataset and comparing the performance of the proposed method with baselines from literature.

5.1 Highlights of Thesis Contributions

The main contributions of this thesis can be summarized as follows:

- **SMC-based privacy preserving federated learning frameworks.** One of the main challenges in federated learning is that training results may expose information about training samples. Chapter 3 addressed this privacy concern of the training results by proposing two different SMC-based frameworks to create secure communication for sharing training results. The first method is proposed for centralized federated learning architecture. In this method, hospitals are divided into small clusters. Hospitals within each cluster collaborate to learn the summation of the local training results without having access to each other's training. The second method is proposed for decentralized federated learning where there is no central server to coordinate training. In this method, training consists of three phases. The first two phases are performed only once at the beginning of the training while the last phase is repeated until either converging or reaching the end of training epochs. The proposed frameworks have been evaluated on kidney slides from the TCGA dataset. The experimental results showed that the proposed methods outperform differential privacy approach.

- **Proportional fair federated learning.** Chapter 4 discussed a novel method, called *proportional fair federated learning* (Prop-FFL), that resolves the challenge of non-IID data distribution among hospitals. The intuition behind the idea is that hospitals cannot equally contribute to the model learning. Hence, their contributions should be *fairly* weighted. The proposed method modifies the optimization objective function at the central server, prevents the model from being biased toward any hospital. Therefore, it encourages the model to have more uniform performance across hospitals. Performance of the proposed method was evaluated on lung and kidney histopathology slides. The experiment results suggested competitive performance with other state-of-the-art methods.

5.1.1 Limitations

The experimental results demonstrated that SMC-based frameworks address the secure communication concern and the proposed fair federated learning encourages uniform performance across hospitals. However, there are some limitations inherent in these methods which should be considered.

- **SMC-based privacy preserving frameworks.** One of the major limitations of the proposed privacy-preserving methods is their communication cost. They maintain privacy, enabling hospitals to gain superior testing accuracy compared to differential privacy at the cost of "communication overhead". However, the medical field is a sensitive domain where stakeholders might be willing to pay reasonable costs in exchange to gain more reliable results.
- **Prop-FFL** The work focused on applying fairness among hospitals in the FedSGD scenario. In this scenario, hospitals need to collaborate with the central server for every batch of data, which leads to communication overhead. However, when participant hospitals have a small dataset, the proposed method would have moderate communication overhead. Small datasets are quite common in the medical data domain as data annotation is expensive and time-consuming. Therefore, when local datasets are small, Prop-FFL can apply fairness among hospitals at acceptable communication costs.

5.2 Future Work

The proposed methods in this thesis open new directions for future work. The main topics will be describe in the following.

5.2.1 Personalized Federated Learning

All frameworks proposed in Chapter 3 and Chapter 4 focus on learning "one single model" for all hospitals. However, in federated learning, each participant might have a different objective. Therefore, allowing each participant to have a private model with any architecture, called *model heterogeneity*, would be beneficial in many applications. Recently, personalized federated learning has gained considerable attention [54, 69]. It may be worth exploring how to apply fairness in personalized federated learning.

5.2.2 Threats and Attacks to Federated Learning

In this thesis, it was assumed that all hospitals collaborate in *good faith*. However, federated learning in a real-world scenario has to be able to detect and confront malicious participants. The participant hospitals might not have malicious intent, however, their local data annotations might be corrupted (or become dorrupted through cyberattacks), comprising the model's performance for others. Various kinds of attacks in federated learning are under-explored in research [70, 71, 72]. A future direction would be to extend Prop-FFL to be stable against attacks.

References

[1] Omar Kujan et al. "Why oral histopathology suffers inter-observer variability on grading oral epithelial dysplasia: an attempt to understand the sources of variation". In: *Oral oncology* 43.3 (2007), pp. 224–231.

[2] Hamid R Tizhoosh et al. "Searching images for consensus: can AI remove observer variability in pathology?" In: *The American journal of pathology* 191.10 (2021), pp. 1702–1708.

[3] Jeroen Van der Laak, Geert Litjens, and Francesco Ciompi. "Deep learning in histopathology: the path to the clinic". In: *Nature medicine* 27.5 (2021), pp. 775–784.

[4] Brian C Drolet et al. "Electronic communication of protected health information: privacy, security, and HIPAA compliance". In: *The Journal of hand surgery* 42.6 (2017), pp. 411–416.

[5] Mark D Zarella et al. "A practical guide to whole slide imaging: a white paper from the digital pathology association". In: *Archives of pathology & laboratory medicine* 143.2 (2019), pp. 222–234.

[6] Muhammad Khalid Khan Niazi, Anil V Parwani, and Metin N Gurcan. "Digital pathology and artificial intelligence". In: *The lancet oncology* 20.5 (2019), e253–e261.

[7] Stephan W Jahn, Markus Plass, and Farid Moinfar. "Digital pathology: advantages, limitations and emerging perspectives". In: *Journal of Clinical Medicine* 9.11 (2020), p. 3697.

[8] Liron Pantanowitz et al. "Whole slide imaging for educational purposes". In: *Journal of pathology informatics* 3.1 (2012), p. 46.

[9] BF Boyce. "Whole slide imaging: uses and limitations for surgical pathology and teaching". In: *Biotechnic & Histochemistry* 90.5 (2015), pp. 321–330.

[10] Gonzalo Romero Lauro et al. “Digital pathology consultations—a new era in digital imaging, challenges and practical applications”. In: *Journal of digital imaging* 26.4 (2013), pp. 668–677.

[11] URL: https://dicom.nema.org/dicom/dicomwsi/,journal={DICOM}.

[12] Markus D Herrmann et al. “Implementing the DICOM standard for digital pathology”. In: *Journal of pathology informatics* 9.1 (2018), p. 37.

[13] David A Clunie. “Dual-Personality DICOM-TIFF for whole slide images: A migration technique for legacy software”. In: *Journal of Pathology Informatics* 10.1 (2019), p. 12.

[14] JD Pallua et al. “The future of pathology is digital”. In: *Pathology-Research and Practice* 216.9 (2020), p. 153040.

[15] Metin N Gurcan et al. “Histopathological image analysis: A review”. In: *IEEE reviews in biomedical engineering* 2 (2009), pp. 147–171.

[16] Hamid Reza Tizhoosh and Liron Pantanowitz. “Artificial intelligence and digital pathology: challenges and opportunities”. In: *Journal of pathology informatics* 9.1 (2018), p. 38.

[17] Andrew Janowczyk and Anant Madabhushi. “Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases”. In: *Journal of pathology informatics* 7.1 (2016), p. 29.

[18] Daisuke Komura and Shumpei Ishikawa. “Machine learning methods for histopathological image analysis”. In: *Computational and structural biotechnology journal* 16 (2018), pp. 34–42.

[19] Shivam Kalra. “Learning Discriminative Representations for Gigapixel Images”. In: (2022).

[20] Maral Rasoolijaberi et al. “Multi-Magnification Image Search in Digital Pathology”. In: *IEEE Journal of Biomedical and Health Informatics* 26.9 (2022), pp. 4611–4622.

[21] Sharyl J Nass, Laura A Levit, Lawrence O Gostin, et al. “The value and importance of health information privacy”. In: *Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research*. National Academies Press (US), 2009.

[22] Nicola Rieke et al. “The future of digital health with federated learning”. In: *NPJ digital medicine* 3.1 (2020), pp. 1–7.

[23] Jia Deng et al. “Imagenet: A large-scale hierarchical image database”. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[24] Georgios A Kaissis et al. “Secure, privacy-preserving and federated machine learning in medical imaging”. In: *Nature Machine Intelligence* 2.6 (2020), pp. 305–311.

[25] Brendan McMahan et al. “Communication-efficient learning of deep networks from decentralized data”. In: *Artificial intelligence and statistics*. PMLR. 2017, pp. 1273–1282.

[26] Tian Li et al. “Federated learning: Challenges, methods, and future directions”. In: *IEEE Signal Processing Magazine* 37.3 (2020), pp. 50–60.

[27] Momina Shaheen et al. “Applications of federated learning; Taxonomy, challenges, and research trends”. In: *Electronics* 11.4 (2022), p. 670.

[28] Chen Zhang et al. “A survey on federated learning”. In: *Knowledge-Based Systems* 216 (2021), p. 106775.

[29] Qiang Yang et al. “Federated learning”. In: *Synthesis Lectures on Artificial Intelligence and Machine Learning* 13.3 (2019), pp. 1–207.

[30] Erfan Darzidehkalani, Mohammad Ghasemi-Rad, and Pma van Ooijen. “Federated Learning in Medical Imaging: Part I: Toward Multicentral Health Care Ecosystems”. In: *Journal of the American College of Radiology* (2022).

[31] Qi Dou et al. “Federated deep learning for detecting COVID-19 lung abnormalities in CT: a privacy-preserving multinational validation study”. In: *NPJ digital medicine* 4.1 (2021), pp. 1–11.

[32] Cosmin Bercea et al. “Federated Disentangled Representation Learning for Unsupervised Brain Anomaly Detection”. In: (2021).

[33] Micah J Sheller et al. “Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data”. In: *Scientific reports* 10.1 (2020), pp. 1–12.

[34] Karthik V Sarma et al. “Federated learning improves site performance in multicenter deep learning without data sharing”. In: *Journal of the American Medical Informatics Association* 28.6 (2021), pp. 1259–1264.

[35] Dong Yang et al. “Federated semi-supervised learning for COVID region segmentation in chest CT using multi-national data from China, Italy, Japan”. In: *Medical image analysis* 70 (2021), p. 101992.

[36] Tariq Bdair, Nassir Navab, and Shadi Albarqouni. “FedPerl: Semi-supervised Peer Learning for Skin Lesion Classification”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2021, pp. 336–346.

[37] Mohammed Adnan et al. “Federated learning and differential privacy for medical image analysis”. In: *Scientific reports* 12.1 (2022), pp. 1–10.

[38] Shivam Kalra et al. “ProxyFL: Decentralized Federated Learning through Proxy Model Sharing”. In: *arXiv preprint arXiv:2111.11343* (2021).

[39] URL: https://www.cancer.gov/tcga.

[40] Ligeng Zhu, Zhijian Liu, and Song Han. “Deep leakage from gradients”. In: *Advances in Neural Information Processing Systems* 32 (2019).

[41] Xuefei Yin, Yanming Zhu, and Jiankun Hu. “A comprehensive survey of privacy-preserving federated learning: A taxonomy, review, and future directions”. In: *ACM Computing Surveys (CSUR)* 54.6 (2021), pp. 1–36.

[42] Martin Abadi et al. “Deep learning with differential privacy”. In: *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. 2016, pp. 308–318.

[43] Hwanjo Yu, Xiaoqian Jiang, and Jaideep Vaidya. “Privacy-preserving SVM using nonlinear kernels on horizontally partitioned data”. In: *Proceedings of the 2006 ACM symposium on Applied computing*. 2006, pp. 603–610.

[44] Yehuda Lindell. “Secure multiparty computation”. In: *Communications of the ACM* 64.1 (2020), pp. 86–96.

[45] Adi Shamir. “How to share a secret”. In: *Communications of the ACM* 22.11 (1979), pp. 612–613.

[46] Alon Brutzkus, Ran Gilad-Bachrach, and Oren Elisha. “Low latency privacy preserving inference”. In: *International Conference on Machine Learning*. PMLR. 2019, pp. 812–821.

[47] Xiaoxiao Li et al. “Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results”. In: *Medical Image Analysis* 65 (2020), p. 101765.

[48] Chuan Zhao et al. “Secure multi-party computation: theory, practice and applications”. In: *Information Sciences* 476 (2019), pp. 357–372.

[49] Yong Li et al. “Privacy-preserving federated learning framework based on chained secure multiparty computing”. In: *IEEE Internet of Things Journal* 8.8 (2020), pp. 6178–6186.

[50] John N Weinstein et al. “The cancer genome atlas pan-cancer analysis project”. In: *Nature genetics* 45.10 (2013), pp. 1113–1120.

[51] Le Hou et al. "Patch-based convolutional neural network for whole slide tissue image classification". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2424–2433.

[52] Gao Huang et al. "Densely connected convolutional networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708.

[53] Maximilian Ilse, Jakub Tomczak, and Max Welling. "Attention-based deep multiple instance learning". In: *International conference on machine learning*. PMLR. 2018, pp. 2127–2136.

[54] Alysa Ziying Tan et al. "Towards personalized federated learning". In: *IEEE Transactions on Neural Networks and Learning Systems* (2022).

[55] Irene Y Chen, Peter Szolovits, and Marzyeh Ghassemi. "Can AI help reduce disparities in general medical and mental health care?" In: *AMA journal of ethics* 21.2 (2019), pp. 167–179.

[56] L Daniel and K Narayanan. "Congestion control 2: Utility, fairness, and optimization in resource allocation". In: *Mathematical Modelling for Computer Networks-Part I* (2013), pp. 2–1.

[57] Tian Li et al. "Fair resource allocation in federated learning". In: *arXiv preprint arXiv:1905.10497* (2019).

[58] Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. "Agnostic federated learning". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 4615–4625.

[59] Tiansheng Huang et al. "An efficiency-boosting client selection scheme for federated learning with fairness guarantee". In: *IEEE Transactions on Parallel and Distributed Systems* 32.7 (2020), pp. 1552–1564.

[60] Lingjuan Lyu et al. "Collaborative fairness in federated learning". In: *Federated Learning*. Springer, 2020, pp. 189–204.

[61] Hoon Kim and Youngnam Han. "A proportional fair scheduling for multicarrier transmission systems". In: *IEEE Communications letters* 9.3 (2005), pp. 210–212.

[62] Li Deng. "The mnist database of handwritten digit images for machine learning research". In: *IEEE Signal Processing Magazine* 29.6 (2012), pp. 141–142.

[63] Han Xiao, Kashif Rasul, and Roland Vollgraf. "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms". In: *arXiv preprint arXiv:1708.07747* (2017).

[64] Milad Sikaroudi, Shahryar Rahnamayan, and HR Tizhoosh. "Hospital-Agnostic Image Representation Learning in Digital Pathology". In: *arXiv preprint arXiv:2204.02404* (2022).

[65] Andrew Janowczyk et al. "HistoQC: an open-source quality control tool for digital pathology slides". In: *JCO clinical cancer informatics* 3 (2019), pp. 1–7.

[66] Juan Antonio Retamero, Jose Aneiros-Fernandez, and Raimundo G Del Moral. "Complete digital pathology for routine histopathology diagnosis in a multicenter hospital network". In: *Archives of pathology & laboratory medicine* 144.2 (2020), pp. 221–228.

[67] Birgid Schömig-Markiefka et al. "Quality control stress test for deep learning-based diagnostic model in digital pathology". In: *Modern Pathology* 34.12 (2021), pp. 2098–2108.

[68] Frederick M Howard et al. "The impact of site-specific digital histology signatures on deep learning model accuracy and bias". In: *Nature communications* 12.1 (2021), pp. 1–13.

[69] Viraj Kulkarni, Milind Kulkarni, and Aniruddha Pant. "Survey of personalization techniques for federated learning". In: *2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)*. IEEE. 2020, pp. 794–797.

[70] Yann Fraboni, Richard Vidal, and Marco Lorenzi. "Free-rider attacks on model aggregation in federated learning". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 1846–1854.

[71] Malhar S Jere, Tyler Farnan, and Farinaz Koushanfar. "A taxonomy of attacks on federated learning". In: *IEEE Security & Privacy* 19.2 (2020), pp. 20–28.

[72] Arjun Nitin Bhagoji et al. "Model poisoning attacks in federated learning". In: *Proc. Workshop Secur. Mach. Learn.(SecML) 32nd Conf. Neural Inf. Process. Syst.(NeurIPS)*. 2018, pp. 1–23.

[73] Muhammad Abdullah Jamal et al. "Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 7610–7619.

[74] Chen-Yu Lee et al. "Sliced wasserstein discrepancy for unsupervised domain adaptation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 10285–10295.

[75] Justin M Johnson and Taghi M Khoshgoftaar. "Survey on deep learning with class imbalance". In: *Journal of Big Data* 6.1 (2019), pp. 1–54.

[76] Shuhan Tan, Xingchao Peng, and Kate Saenko. "Class-Imbalanced Domain Adaptation: An Empirical Odyssey". In: *European Conference on Computer Vision*. Springer. 2020, pp. 585–602.

[77] Tsung-Yi Lin et al. "Focal loss for dense object detection". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2980–2988.

[78] Han Zhao et al. "On learning invariant representations for domain adaptation". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 7523–7532.

[79] MICCAI 2020 Challenge PANDA Dataset. Accessed June. 10, 2021. MICCAI 2020 Challenge. URL: https://www.kaggle.com/c/prostate-cancer-grade-assessment/data.

[80] Jian Ren et al. "Adversarial domain adaptation for classification of prostate histopathology whole-slide images". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 201–209.

[81] Yaroslav Ganin and Victor Lempitsky. "Unsupervised domain adaptation by backpropagation". In: *International conference on machine learning*. PMLR. 2015, pp. 1180–1189.

[82] Iaroslav Melekhov, Juho Kannala, and Esa Rahtu. "Siamese network features for image matching". In: *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE. 2016, pp. 378–383.

[83] Nitesh V Chawla, Nathalie Japkowicz, and Aleksander Kotcz. "Special issue on learning from imbalanced data sets". In: *ACM SIGKDD explorations newsletter* 6.1 (2004), pp. 1–6.

[84] Jason Van Hulse, Taghi M Khoshgoftaar, and Amri Napolitano. "Experimental perspectives on learning from imbalanced data". In: *Proceedings of the 24th international conference on Machine learning*. 2007, pp. 935–942.

[85] Abolfazl Farahani et al. "A brief review of domain adaptation". In: *arXiv preprint arXiv:2010.03978* (2020).

[86] Ian Goodfellow et al. "Generative adversarial nets". In: *Advances in neural information processing systems* 27 (2014).

[87] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. "Domain adaptation for large-scale sentiment classification: A deep learning approach". In: *ICML*. 2011.

[88] Yeganeh Madadi et al. “Deep visual unsupervised domain adaptation for classification tasks: a survey”. In: *IET Image Processing* (2020).

[89] Kaichao You et al. “Universal domain adaptation”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 2720–2729.

[90] Pau Panareda Busto and Juergen Gall. “Open set domain adaptation”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 754–763.

[91] Zhangjie Cao et al. “Partial transfer learning with selective adversarial networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 2724–2732.

[92] Fuzhen Zhuang et al. “A comprehensive survey on transfer learning”. In: *Proceedings of the IEEE* 109.1 (2020), pp. 43–76.

[93] Arthur Gretton et al. “A kernel two-sample test”. In: *The Journal of Machine Learning Research* 13.1 (2012), pp. 723–773.

[94] Baochen Sun, Jiashi Feng, and Kate Saenko. “Return of frustratingly easy domain adaptation”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 30. 1. 2016.

[95] Baochen Sun and Kate Saenko. “Deep coral: Correlation alignment for deep domain adaptation”. In: *European conference on computer vision*. Springer. 2016, pp. 443–450.

[96] Sergey Ioffe and Christian Szegedy. “Batch normalization: Accelerating deep network training by reducing internal covariate shift”. In: *International conference on machine learning*. PMLR. 2015, pp. 448–456.

[97] Yujia Li, Kevin Swersky, and Rich Zemel. “Generative moment matching networks”. In: *International Conference on Machine Learning*. PMLR. 2015, pp. 1718–1727.

[98] Xingchao Peng et al. “Moment matching for multi-source domain adaptation”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 1406–1415.

[99] Yves Grandvalet, Yoshua Bengio, et al. “Semi-supervised learning by entropy minimization.” In: *CAP* 367 (2005), pp. 281–296.

[100] Mei Wang and Weihong Deng. “Deep visual domain adaptation: A survey”. In: *Neurocomputing* 312 (2018), pp. 135–153.

[101] Konstantinos Bousmalis et al. “Domain separation networks”. In: *Advances in neural information processing systems* 29 (2016), pp. 343–351.

[102] Guy Nir et al. “Automatic grading of prostate cancer in digitized histopathology images: Learning from multiple experts”. In: *Medical image analysis* 50 (2018), pp. 167–180.

[103] Eirini Arvaniti et al. “Automated Gleason grading of prostate cancer tissue microarrays via deep learning”. In: *Scientific reports* 8.1 (2018), pp. 1–11.

[104] Zhengming Ding, Ming Shao, and Yun Fu. “Deep low-rank coding for transfer learning”. In: *Twenty-Fourth International Joint Conference on Artificial Intelligence*. 2015.

[105] “Federated Learning: Collaborative Machine Learning without Centralized Training Data”. In: *https://ai.googleblog.com/2017/04/federated-learning-collaborative.html*.

[106] Nguyen H Tran et al. “Federated learning over wireless networks: Optimization model design and analysis”. In: *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE. 2019, pp. 1387–1395.

[107] Quande Liu et al. “Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 1013–1023.

[108] “FedJAX: Federated Learning Simulation with JAX”. In: URL: https://ai.googleblog.com/2021/10/fedjax-federated-learning-simulation.html.

[109] Tomer Gafni et al. “Federated learning: A signal processing perspective”. In: *arXiv preprint arXiv:2103.17150* (2021).

[110] Daniel Peterson, Pallika Kanani, and Virendra J Marathe. “Private federated learning with domain adaptation”. In: *arXiv preprint arXiv:1912.06733* (2019).

[111] “What Is Federated Learning?” In: *https://blogs.nvidia.com/blog/2019/10/13/what-is-federated-learning/*.

[112] Bartosz Krawczyk. “Learning from imbalanced data: open challenges and future directions”. In: *Progress in Artificial Intelligence* 5.4 (2016), pp. 221–232.

[113] “Prostate Cancer Grading & Prognostic Scoring”. In: *https://www.prostateconditions.org/about-prostate-conditions/prostate-cancer/newly-diagnosed/gleason-score*.

[114] “Radboud University Medical Center”. In: *https://www.radboudumc.nl/en/research*.

[115] “Karolinska Institute”. In: *https://ki.se/en/meb*.

[116] Jonathan I Epstein et al. “The 2014 International Society of Urological Pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma”. In: *The American journal of surgical pathology* 40.2 (2016), pp. 244–252.

[117] Phillip M Pierorazio et al. “Prognostic Gleason grade grouping: data based on the modified Gleason scoring system”. In: *BJU international* 111.5 (2013), p. 753.

[118] Kyle Chang et al. “The cancer genome atlas pan-cancer analysis project”. In: *Nat Genet* 45.10 (2013), pp. 1113–1120.

[119] Miriam Hägele et al. “Resolving challenges in deep learning-based analyses of histopathological images using explanation methods”. In: *Scientific reports* 10.1 (2020), pp. 1–12.

[120] Robert Geirhos et al. “Shortcut learning in deep neural networks”. In: *Nature Machine Intelligence* 2.11 (2020), pp. 665–673.

[121] Da Li et al. “Deeper, broader and artier domain generalization”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 5542–5550.

[122] Mehmet Sezgin and Bülent Sankur. “Survey over image thresholding techniques and quantitative performance evaluation”. In: *Journal of Electronic imaging* 13.1 (2004), pp. 146–165.

[123] Nobuyuki Otsu. “A threshold selection method from gray-level histograms”. In: *IEEE transactions on systems, man, and cybernetics* 9.1 (1979), pp. 62–66.

[124] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

APPENDICES

Appendix A

Proportional Fairness

Proof. The optimization problem (4.4) is concave with a closed form optimal solution [124]. Using product rule of logarithms, we can simplify (4.4) to following

$$\max_{w} \sum_{k=1}^{M} \log\big(p_k\big) + \log\big(F'_k(w)\big),$$
$$\text{s.t.} \quad F'_k(w) = \frac{F_k(w)}{\sum_{j=1}^{M} F_j(w)}.$$

Since p_k is not a function of w, we can remove it from the objective function. Also, we replace the constraint above with the hidden constraint $\sum_{k=1}^{M} F'_k(w) = 1$. Therefore, the optimization problem will simplified to

$$\max_{w} \sum_{k=1}^{M} \log\big(F'_k(w)\big), \quad \text{s.t.} \quad \sum_{k=1}^{M} F'_k(w) = 1.$$

We solve this optimization problem by employing Lagrangian approach. Let μ be the Lagrangian multiplier. Then, we can convert the optimization problem to

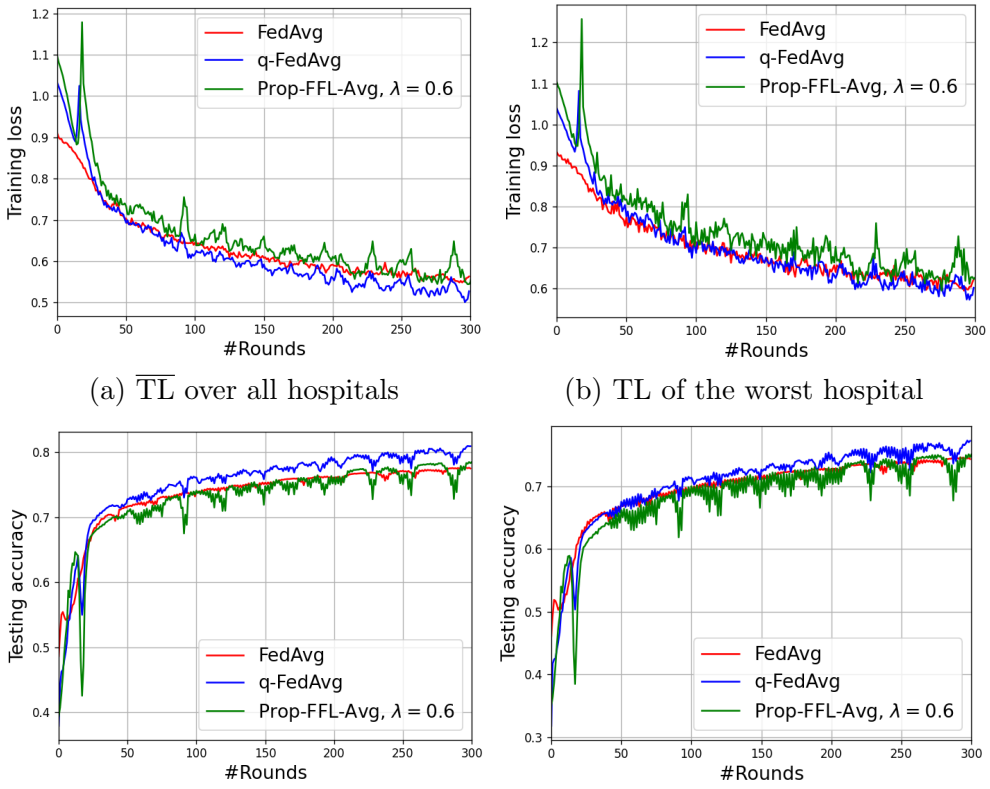
$$\min_{w,\mu} -\sum_{k=1}^{M} \log\big(F'_k(w)\big) + \mu\big(\sum_{k=1}^{M} F'_k(w) - 1\big).$$

Using first order stationary condition, we get $\frac{1}{F'^{*}_k(w)} = \mu$ and $\sum_{k=1}^{M} F'^{*}_k(w) = 1$. This gives $\mu = M$ and $F'^{*}_k(w) = \frac{1}{M}$, which is an optimal solution that satisfies all constraints and KKT conditions.

Appendix B

Expansion of Prop-FFL on FedAvg

In this section, we validate the performance of the proposed approach on top of the FedAVG. It means we allow hospitals to train the model on their local data multiple times, updating the model repeatedly before sending training results to the central server. The model parameters including λ, q, and learning rate have been tuned for each dataset to get the best possible performance in each of those three methods. The experimental results have been presented for both histopathology datasets in Fig. B.1 and B.2. As can be seen in these figures, the results of Prop-FFL are not promising compared to the other two methods. This happens because the fairness loss term is only considered in the central server aggregation and its impact is considerably reduced in FedAVG. However, we believe that modifying local training at each hospital by changing the local loss function can make Prop-FFL suitable for the FedAVG scenario too. This will be left for the future works.



(a) $\overline{\text{TL}}$ over all hospitals

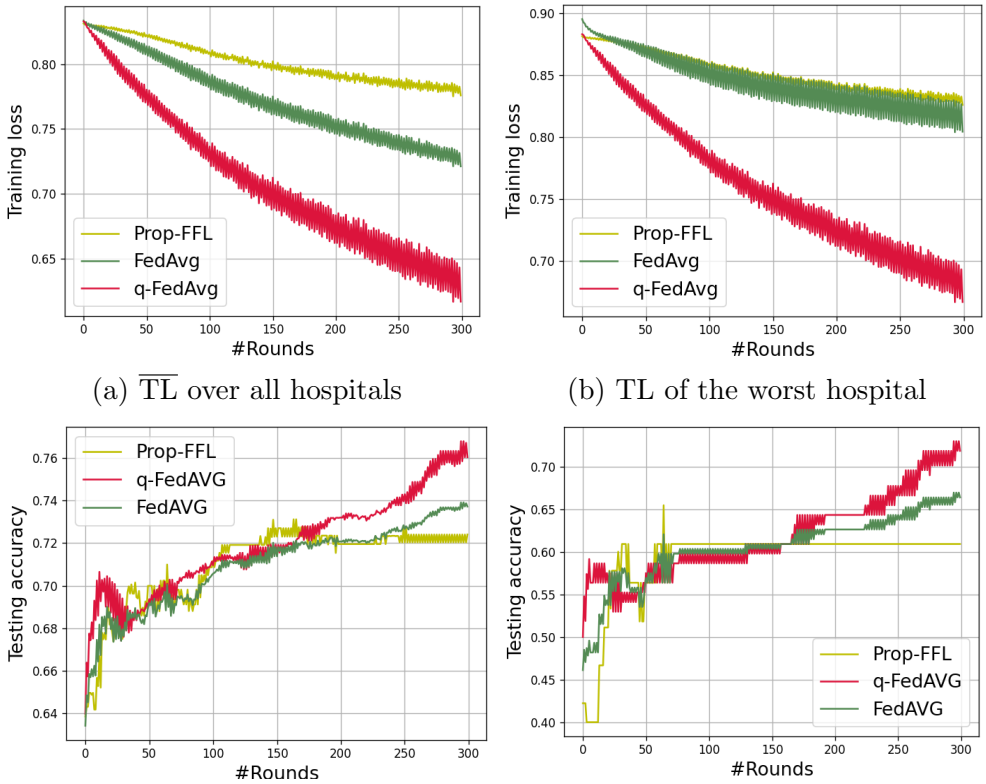
(b) TL of the worst hospital

(c) $\overline{\text{TA}}$ of all hospitals

(d) TA of the worst hospital

Figure B.1: Evaluation in FedAvg scenario on **Histopathology-Kidney** dataset. The results for Prop-FFL have been provided for default $\lambda = 0.6$. (TL= training loss, $\overline{\text{TL}}$=average training loss), TA= testing accuracy, $\overline{\text{TA}}$=average testing accuracy)

.



(a) $\overline{\text{TL}}$ over all hospitals

(b) TL of the worst hospital

(c) $\overline{\text{TA}}$ of all hospitals

(d) TA of the worst hospital

Figure B.2: Evaluation in FedAvg scenario on **Histopathology Lung** dataset. The results for Prop-FFL have been provided for default $\lambda = 0.6$. (TL= training loss, $\overline{\text{TL}}$=average training loss), TA= testing accuracy, $\overline{\text{TA}}$=average testing accuracy)