Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

School of Computer Science and Statistics

Application of Personalised Federated Learning for Pneumonia Detection in Chest X-Rays

Tom Roberts

Supervisor: Meriel Huggard

April 22nd, 2024

A dissertation submitted in partial fulfilment
of the requirements for the degree of
Masters in Computer Science

Declaration

I hereby declare that this dissertation is entirely my own work and that it has not been submitted as an exercise for a degree at this or any other university.

I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at `http://www.tcd.ie/calendar`.

I have completed the Online Tutorial on avoiding plagiarism 'Ready Steady Write', located at `http://tcd-ie.libguides.com/plagiarism/ready-steady-write`.

I consent / do not consent to the examiner retaining a copy of the thesis beyond the examining period, should they so wish (EU GDPR May 2018).

I agree that this thesis will not be publicly available, but will be available to TCD staff and students in the University's open access institutional repository on the Trinity domain only, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement. **Please consult with your supervisor on this last item before agreeing, and delete if you do not consent**

Signed: _____ Date: _____

Tom Roberts April 15th, 2024

Abstract

This dissertation aims to explore the application of personalised federated learning for the detection of pneumonia in chest X-ray images. The primary goal of this project is to investigate the challenges posed by non-IID data across numerous distributed datasets, and to explore techniques to mitigate the effects of non-IID data.

Making use of publicly available datasets, a federated learning model for the detection of pneumonia in chest X-ray images is created. Techniques inspired by the state-of-the-art are then implemented, to mitigate the effects of non-IID data on the performance of the models. The findings demonstrate the importance of personalisation techniques in improving the performance of federated learning models, particularly in the context of medical image classification.

Lay Abstract

This research project investigates a unique approach to machine learning, in order to improve the accuracy of models used for detecting pneumonia in chest X-ray images by using a method called personalised federated learning. Unlike traditional machine learning methods, which rely on combining all of the data in a central location, federated learning uses data which is scattered across different locations. This is to ensure that the data remains private and secure.

However, since the data in different locations can have different characteristics, a "one-size-fits-all" model often fails to perform well, as it is unable to adapt to the differences in the data, This study explores and evaluates a number of techniques that can be used to improve the performance of these models.

The results show that models which implement these "personalisation" techniques are able to detect pneumonia more accurately than models which do not. The results of this study show that personalised federated learning can be an effective approach for improving the accuracy of machine learning models in medical applications.

Acknowledgements

Firstly, I would like to thank my supervisor, Dr. Meriel Huggard, for her guidance and support throughout the course of this project.

I would also like to thank my family and friends for their support, and for keeping me sane over the course of the year.

Finally, I would like to thank the School of Computer Science and Statistics for providing me with the opportunity to undertake this project.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Federated Learning

Machine learning is a subset of artificial intelligence which focuses on the development of algorithms, which can be used in a wide range of applications. These algorithms, commonly known as models, can learn from and make predictions or decisions based on data it has been trained on.

Convolutional Neural Networks (CNN) are a type of machine learning model that are commonly used for a number of complex applications, including image recognition, natural language processing, and speech recognition. CNNs have been shown to be very effective in these applications, achieving record breaking results on highly-challenging datasets such as ImageNet [1]. These models require to be trained on large datasets, where the input data is passed through a series of layers, which extract features from the data and make predictions based on these features.

Traditionally, these models must be trained on a centralised server, which requires the data to be sent to the server for training. This can cause a number of privacy concerns regarding the data. This is particularly true in the case of medical data, where data is often highly sensitive and protected by law in many jurisdictions such as the *General Data Protection Regulation (GDPR)* [2] in the European Union and the *Health Insurance Portability and Accountability Act (HIPAA)* [3] in the United States.

To combat the privacy concerns of traditional machine learning models, Federated Learning (FL) was conceived by Google in 2017 [4]. FL is a decentralised approach to machine learning which allows models to be trained on data that is distributed across a number of devices. This approach allows the data to remain on local devices, and only the model updates are sent to a centralised server. This alleviates the privacy concerns associated with sending sensitive data to a centralised server.

One practical example of federated learning that has been implemented is the *Gboard*

Keyboard [5] on Android devices, which uses federated learning to improve the text prediction feature. The model is trained locally on the device, only sending model updates to the server. This allows the model to be trained on the user's personal typing habits, without the concern of sensitive data being sent to a central server.

Another useful application of federated learning is in the healthcare industry, where data is often highly sensitive and often subject to strict privacy laws.

1.2 Problem Statement

Convolutional neural networks rely on an algorithm called Stochastic Gradient Descent (SGD) to train and update the weights of the model. However, since federated learning trains a number of separate models independently, the model weights will differ on each device. This can lead to a statistical bias in the local models, which can result in the global model performing poorly. This is particularly true in the case of medical data, where the data is highly prone to non-IID trends, due to the variation in patient demographics, image quality, and data collection procedures.

Non-IID (non identically and independently distributed) data refers to data that is not uniformly distributed across all devices, often resulting in data samples which are not representative of the global data distribution, and which vary significantly in terms of features, labels, or distributions. This can lead to models which are biased towards the local data, and which struggle to generalise to the global data distribution [6]. This can result in models making incorrect predictions, which can have serious consequences in a clinical setting.

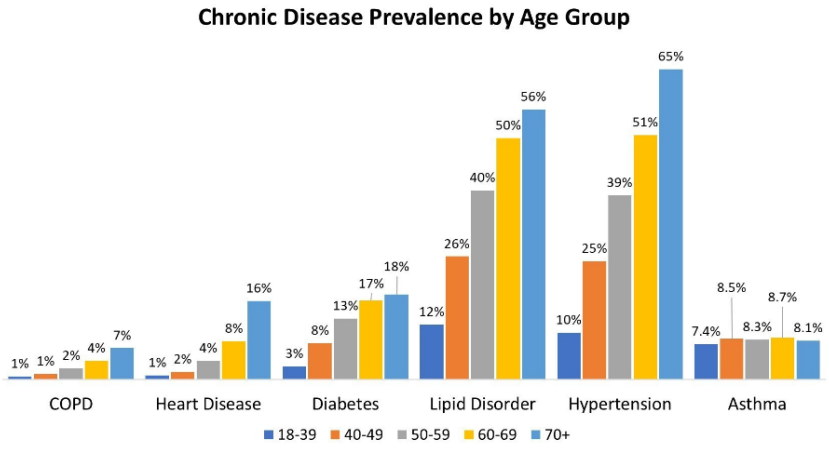

Figure 1.1: Chronic Disease Prevalence by Age Group - Patient data provided from US healthcare provider Catapult Health [7]. The first step involves identifying the body parts in the X-ray image, while the second step involves classifying these parts as either normal or abnormal.

An example of this can be seen in Figure 1.1, which shows the prevalence of chronic disease by age group. Showing that the prevalence of chronic disease increases with age, with the highest prevalence in patients over the age of 60. This trend is not unique to chronic diseases, and can be seen in a number of different medical conditions. This poses a challenge for machine learning models, as the models must be able to adapt to these non-IID trends in the data, without compromising the learning of other trends.

The goal of this project is to investigate the impact of non-IID data on the performance of federated learning models, and to explore techniques to mitigate the effects of non-IID data. The project will focus specifically on the detection of pneumonia in chest X-ray images, a common medical imaging task which has been widely studied in the context of machine learning. The project will explore the challenges posed by non-IID data across locally distributed datasets, and will investigate techniques to adapt the models to the non-IID trends in the data.

1.2.1 Motivation & Research Objectives

Since the introduction of federated learning, there has been numerous efforts made to solve the issues that arise from non-IID data. These efforts include the adjustment of federated learning algorithms, the development of new algorithms, and the use of personalisation techniques. Algorithmic adjustments aim to improve the performance of the global model by counteracting the effects of non-IID data on local models. Personalisation techniques aim to adapt the models to the non-IID trends in the data, without compromising the learning of

other trends.

In this paper, we aim to investigate the following research objectives:

- Investigate the impact of non-IID data on the performance of federated learning models, specifically in the context of medical image classification.
- Review the current state-of-the-art in federated learning, and techniques used to mitigate the effects of non-IID data.
- Develop a federated learning model for the detection of pneumonia in chest X-ray images.
- Implement an approach, inspired by the state-of-the-art, to mitigate the effects of non-IID data on the performance of the models.
- Using metrics, evaluate the success of this approach by comparing the performance of the global model to the personalised models.

The next chapter will provide a review of the current state-of-the-art in federated learning, and techniques used to mitigate the effects of non-IID data, followed by relevant examples of machine learning in the medical field.

2 Background & Literature Review

2.1 Federated Learning

Federated Learning (FL) is a decentralised approach to machine learning, where the goal is to train a model across a number of decentralised devices that hold local data samples, without the need to need to send or store the data on a centralised server. The concept of federated learning was first introduced by McMahan et al. in 2017 [4], where the authors proposed the method, and demonstrated it's effectiveness in training deep networks from decentralised data.

One of the primary benefits of federated learning is it's ability to enhance user privacy and data security. By design, any sensitive data remains on the local device, and only the model updates are sent to the central server. Therefore, it complies with data protection laws which prohibit the sharing of sensitive data, such as the *General Data Protection Regulation (GDPR)* [2] in the European Union and the *Health Insurance Portability and Accountability Act (HIPAA)* [3] in the United States.

Federated learning inherently involves a number of challenges, most notably the issue of dealing with unbalanced and non-IID (non-independently and identically distributed) data. These challenges introduce the need for new strategies and algorithms to be developed to address these issues, which will be discussed in the sections following.

The general federated learning process is as follows:

1. **Initialise Global Model:** The global model is initialised on the central server, which is used as a starting point for the training process. This model is then distributed to the local devices.
2. **Local Model Training:** Each model participating in the federated learning process trains the model locally using it's own data. This training process typically involves the use of Stochastic Gradient Descent (SGD), or a similar algorithm to update the

model weights.

3. **Aggregation of Model Updates:** The model updates from each device are then sent to the central server, where they are aggregated to form a new global model. This aggregation process can be done in a number of ways, using algorithms such as Federated Averaging, or FedSGD [4].

4. **Model Update:** The new set of global model parameters are then sent back to the local devices, where each device will update it's local model with the new parameters. This process of training, aggregation and updating is repeated iteratively until the model converges, or until a stopping criterion is met.

2.1.1 Federated Averaging, FedSGD

Federated Averaging, also known as FedAvg, is a popular algorithm used in federated learning to aggregate model updates from multiple devices. This algorithm was first introduced along with the concept of federated learning by McMahan et al. in 2017 [4].

Algorithm 1 `FederatedAveraging`. The K clients are indexed by k; B is the local minibatch size, E is the number of local epochs, and η is the learning rate.

Server executes:
initialize w_0
for each round $t = 1, 2, \dots$ **do**
  $m \leftarrow \max(C \cdot K, 1)$
  $S_t \leftarrow$ (random set of m clients)
  for each client $k \in S_t$ **in parallel do**
    $w_{t+1}^k \leftarrow \text{ClientUpdate}(k, w_t)$
  $m_t \leftarrow \sum_{k \in S_t} n_k$
  $w_{t+1} \leftarrow \sum_{k \in S_t} \frac{n_k}{m_t} w_{t+1}^k$ *// Erratum*[4]

ClientUpdate(k, w): *// Run on client k*
$\mathcal{B} \leftarrow$ (split $\mathcal{P}_k$ into batches of size B)
for each local epoch i from 1 to E **do**
  for batch $b \in \mathcal{B}$ **do**
    $w \leftarrow w - \eta \nabla \ell(w; b)$
return w to server

Figure 2.1: Federated Averaging Algorithm [4]

Federated averaging is used to aggregate the model updates from each device, then calculate a weighted average of the model updates to form a new global model. The specific algorithms used by the central server and distributed client can be seen in Figure 2.1.

While the goal of federated averaging is to reduce the statistical bias in the model updates, it still has a number of limitations. One of the main limitations is that it assumes that the data between each device is identically and independently distributed (IID). However, in practice, the data distribution between devices is often non-IID, which can lead to a number of issues, such as poor model performance and slow convergence. This issue will be discussed in more detail in the following sections.

2.2 Non-IID Data

Non-IID data refers to data which is not identical and independently distributed across devices. In the context of federated learning, this means that the local data samples on each device are not representative of the entire dataset. This can be caused by a number of reasons, such as population/geographical differences, or the way in which the data was collected.

2.2.1 Effects of Non-IID Data

The vast majority of data distributions in real-world scenarios are non-IID. This is not usually an issue for centralized machine learning models because all of the data is stored in one location, allowing the model to be trained on the entire dataset. However, in the case of federated learning, the data is distributed across multiple devices, and the model must be trained on each device independently. The decentralised nature of federated learning can make non-IID trends more apparent, which can lead to a number of issues.

Federated learning models trained on non-IID data often result in poor model performance, negatively affecting both the accuracy and the convergence speed [8]. If the local datasets vary significantly, the model updates from each client will essentially push the global model in different directions. This can cause the model weights to diverge, rather then converge, as each round of aggregation will try to find a compromise between the conflicting updates. Figure 2.2 shows the effect of non-IID data on federated learning, where the model weights diverge. This divergence of model weights is often amplified as more rounds of training are performed.

The specific challenges caused by non-IID data in federated learning introduce the need for new strategies and algorithms to address these issues. The following sections will discuss some of the current approaches to solving the issue of non-IID data in federated learning.

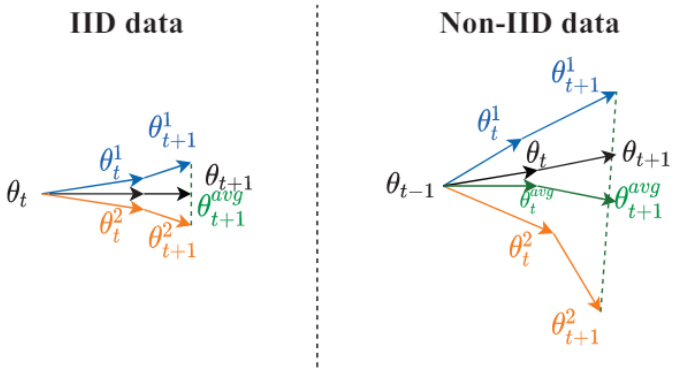

Figure 2.2: Effect of Non-IID Data on Federated Learning [9].
Where θ_t represents the global model and θ_t^{avg} is the averaged model of the local client models.

2.3 Solving the issue of Non-IID Data

Having previously discussed the challenges of non-IID data in federated learning, it is now time to review the state-of-the-art solutions to address these challenges.

2.3.1 Algorithmic Solutions

FedProx is an algorithm which has been designed to address some of the challenges of non-IID data in federated learning. FedProx is a modification of the FedAvg algorithm, which introduces a proximal term in order to account for the model parameter shift across clients [10].

The proximal term acts as a regularisation mechanism, penalising large deviations in model updates which encourages the local model parameters to stay closer to the global model. FedProx is often more suitable than FedAvg in situations where the data is non-IID. The proximal term allows FedProx to converge faster by limiting the divergence of model weights. It also allows the model to maintain better generalisation across diverse clients.

However, by constraining the model updates to not deviate too far from the global model, the local model may not be able to accurately fit local data which is significantly different from the global model. FedProx can also slightly increase the computational cost of training, as the proximal term requires additional computation. Overall FedProx addresses some of the challenges of non-IID data in federated learning, and has been shown to outperform FedAvg in a number of scenarios [10].

Another algorithm which has been designed to combat the effects of non-IID data in federated learning is stochastic controlled averaging (*SCAFFOLD*). SCAFFOLD makes use

of control variates to reduce the variance, and correct the bias which can occur when updating local models [11]. This algorithm maintains a server-side control variate, along with client specific control variates, which are used to correct the bias in the model updates with the aim to align them more closely with the global model. Figure 2.4 shows the update steps of SCAFFOLD on an individual client, highlighting the correction term which ensures that the model updates move towards the global optimum.

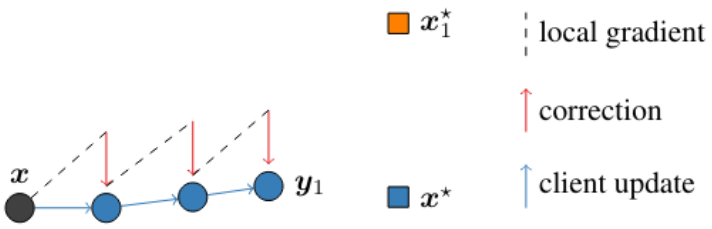

Figure 2.3: Update steps of SCAFFOLD on an individual client [11]. The local model gradient is represented by the dashed black line, and points to x_1^*. The correction term ensures that the model updates move towards the global optimum x^*. The correction term is defined as the difference between the server control variate and the local control variate, and is represented by the dashed red line.

SCAFFOLD effectively counters the effects of non-IID data by reducing the variance in the model updates, and ensuring that the model updates are more aligned with the global model. This allows SCAFFOLD to converge faster, and results in a more accurate model than FedAvg, particularly when the data is highly diverse across clients. Overall, SCAFFOLD effectively combats several of the challenges of non-IID data in federated learning, reducing the model update variance, ensuring that the model updates are more aligned with the global model. [11].

2.3.2 Personalisation

The algorithmic solutions discussed in the previous section have been shown to be effective at combatting the effects of non-IID data in federated learning. However, these solutions treat non-IID data as a problem to be solved, rather than developing a method to leverage the diversity of the data. This is where personalisation come into play.

Personalisation in federated learning refers to the process of training a model on a diverse set of data, and then adapting each local model to the specific characteristics of the local data. This allows each local model to adapt to the unique features of the local data, while still benefiting from the global model. Essentially, personalisation consists of any techniques which exploit non-IID data to improve the performance of the model.

One of the most common methods of personalisation is to use *transfer learning*. Transfer learning is a machine learning technique where a model which was trained for a particular task is reused as the starting point for a new model. One of the main advantages of transfer learning is that it allows a model which has been trained on a large dataset to be fine-tuned on a much smaller, local dataset. One specific challenge that arises with federated transfer learning is the issue of negative transfer, where the knowledge from the global model actually hinders the performance of the local model. This can occur when the global model is not well suited to the local data, and can cause the local model to perform poorly. One way to address this issue would be to implement a mechanism which maximises the positive impact of personalisation, while minimising the risks of negative transfer. [12]

Another method of personalisation is to use *multi-task learning*. Multi-task learning (MTL) is a machine learning technique where a model is trained to perform multiple tasks simultaneously. This can be applied to federated learning, which allows for the simultaneous training of multiple models on different tasks. By sharing parameters across tasks, MTL reduces redundant learning, while also improving the generalisation of the model [13].

2.4 Examples of Machine Learning in Healthcare

2.4.1 Two-step X-ray Image Classification

In the context of medical imaging, classifying X-ray images can serve as a useful tool for assisting radiological diagnosis, especially in situations where there is a shortage of radiologists. The paper "X-ray Image Classification Using Two-step DenseNet Classifiers" by Gomes and Lawal [14] propose an advancement in the classification of X-ray images by using a two-step classification process. This process uses DenseNet [15] classifiers are designed to improve the accuracy of predictions. This two-step method addresses the identification of particular body parts, then uses this information to classify these parts as either normal or abnormal.

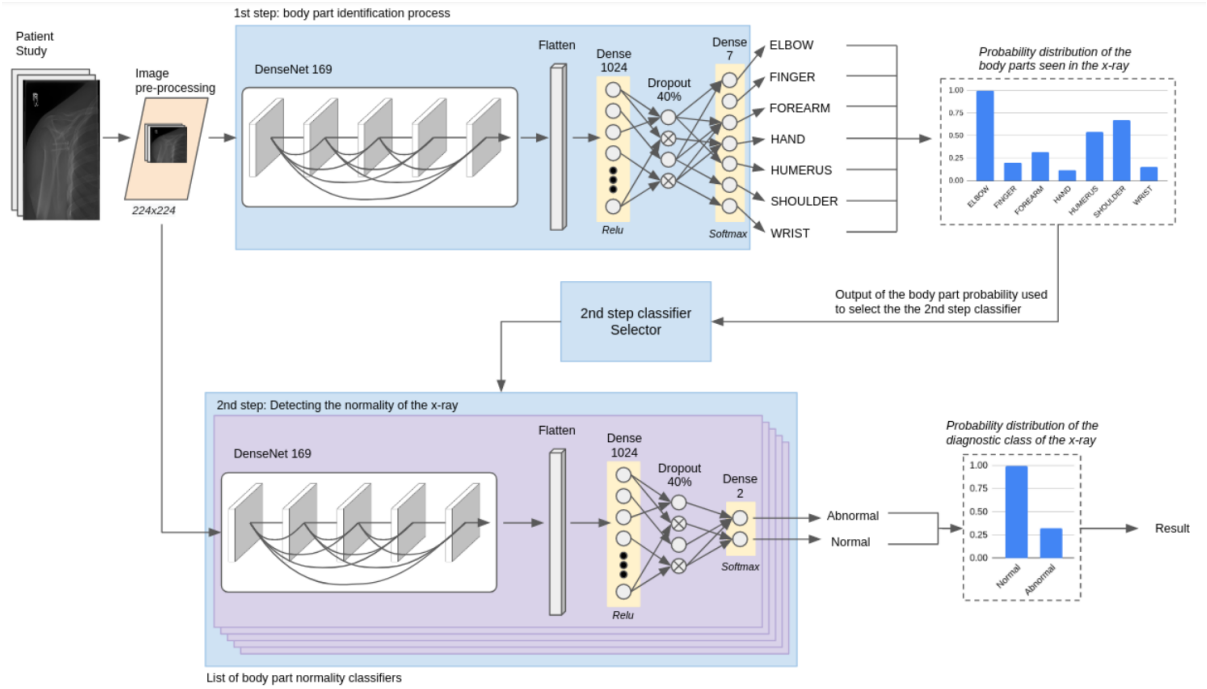

Figure 2.4: The two-step classification process proposed by Gomes and Lawal [14]. The first step involves identifying the body parts in the X-ray image, while the second step involves classifying these parts as either normal or abnormal.

The method can be described as a "divide-and-conquer" strategy, where the images are first categorised based on the detected body parts, before being classified as normal or abnormal. This approach allows for the model to focus on specific features in the image, then uses this information to classify the image. This particular method could be useful in the context of federated learning, as it allows for the model to easily adapt to specific features in the data. This could be particularly useful in the context of non-IID data. For example, when detecting the presence of pneumonia in the lungs, the model could first identify the lungs in the image, then use this information to classify the lungs as either normal or abnormal. This would allow the model to adapt to the specific features of the data, while also ignoring any abnormalities that could be present in other parts of the image.

2.4.2 Personalized Federated Learning: In-Home Health Monitoring

In the paper titled "FedHome: Personalized Federated Learning for In-Home Health Monitoring" by Li et al. [16], a method for personalized federated learning is proposed for in-home health monitoring. As the population ages, the demand for in-home health monitoring systems is increasing. However, the data collected from these systems is highly sensitive, the users health data can not be shared with a centralised server. This is where federated learning comes into play, as it allows for the model to be trained on the local

device, without the need to send the data to a central server.

By combining federated learning with cloud-edge computing architecture, the authors aim to develop an advanced health monitoring system which is not only privacy-preserving, but also personalised to the individual user.

FedHome is a personalised federated learning system which is based on cloud-edge computing architecture. The system consists of a central server, which is responsible for aggregating the model updates from the local devices, along with a number of edge devices, which are used to train the local models. The system architecture is shown in Figure 2.5.

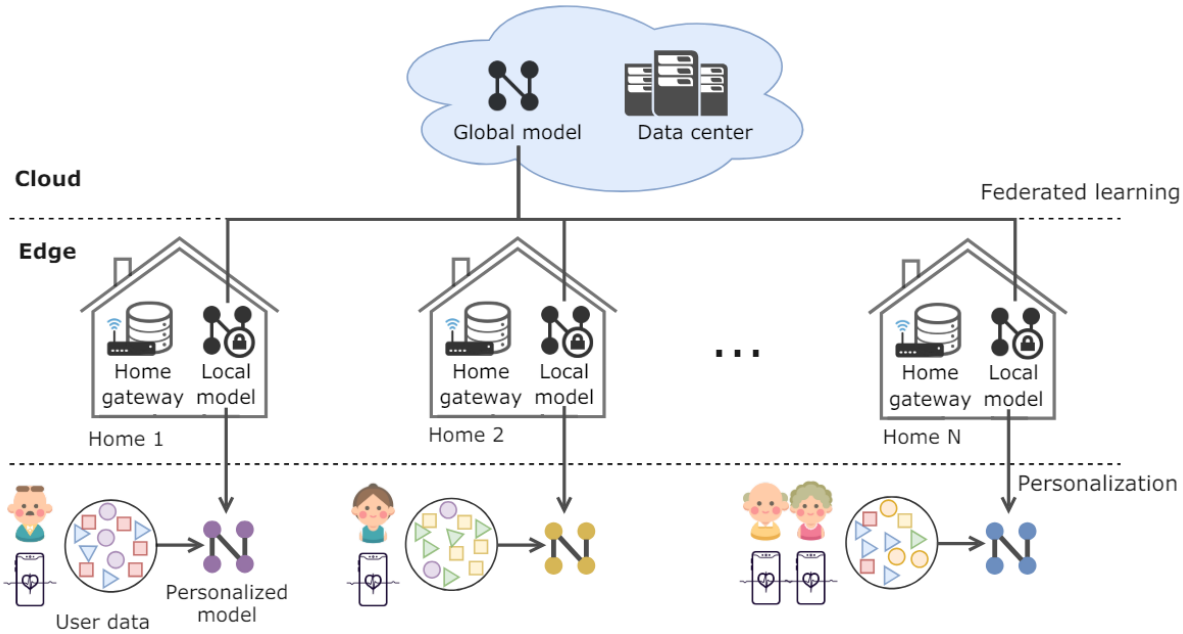

Figure 2.5: Personalisation Strategy proposed by FedHome [16], where multiple homes participate in the federated learning process produce a global model. Then, a local model is generated using the global model, and the user's local data. Using their own data, a user can further personalise their data by generating synthetic data, which is used to further refine the local model parameters.

To deal with the issue of non-IID data, FedHome implements a very unique personalisation strategy. FedHome makes use of a generative convolutional encoder (GCAE) [17], which is used to generate synthetic data based on the data of the user, which is generated using a synthetic minority over-sampling technique (SMOTE) [18]. This synthetic data is then used to fine-tune the model, allowing the model to adapt to the specific features of the user's data, without transmitting any sensitive data. The user can choose whether they want to use personalised training for themselves, or a clustered personalisation strategy, which groups users with similar data together. Both of these strategies involve adapting their personal models to the synthetic data. The main weakness to this approach is that the system is

computationally intensive, as it requires the generation of synthetic data for each user. There are also many hyperparameters, which can be difficult to tune for optimal performance.

In the paper, a number of extensive experiments were conducted to evaluate the performance of FedHome. The results showed FedHome outperformed traditional centralised learning methods, for both balanced and unbalanced datasets. FedHome achieved an accuracy of 95.41%, with over a 7.48% improvement over convolutional neural networks (CNNs) trained on the same data. This demonstrates the effectiveness of personalised federated learning in the context of in-home health monitoring, and the potential for federated learning to be used in a wide range of applications.

3 Technical Content & Project Execution

3.1 Overview of Methodology

The aim of this project is to identify non-IID trends which are present in smart healthcare datasets, and to develop a federated learning system which can effectively counteract the negative effects of these trends.

To achieve this, I decided to focus on the detection of pneumonia in chest x-ray images, using a federated learning system which is designed to work with non-IID data. Pneumonia was chosen as the target disease, as it is a common and serious illness which can be detected using chest x-ray images. Especially in recent times, with the outbreak of the COVID-19 pandemic, the detection of pneumonia has become increasingly important.

Due to the nature of the data involved with chest x-ray images, there is a high probability that it will be non-IID. This is because the data is very likely to have been collected from a number of different sources, such as different hospitals, which each may have different imaging equipment, different patient demographics, and different imaging protocols. This makes it ideal candidate for testing the effectiveness of a federated learning system which is designed to combat the effects of non-IID data.

The implementation of a personalisation layer was chosen as the method to counteract the non-IID trends in the data. This layer is designed to allow each device to train a personalised model, which is then used to make predictions on the local data. The model updates are then sent to the central server, where they are aggregated and used to update the global model. This approach is designed to allow the model to learn from the local data, while still benefiting from the knowledge provided by the global model.

In this chapter, we will:

- Outline the datasets used in the project, and the pre-processing steps which were taken to prepare the data for training.
- Identify examples of non-IID trends in the data, and discuss the implications of these trends on the performance of the model.
- Describe the personalisation strategy which was implemented, and the methods which were used to train the personalised models.

3.2 Pneumonia Detection with Federated Learning

3.2.1 Datasets & Data Pre-Processing

The objective of this project is to develop a federated learning system which can effectively detect pneumonia in chest x-ray images. To achieve this, a number of different datasets were used, which contained chest x-ray images of patients with and without pneumonia. Below is a list of the datasets which were used in this project.

- **CXRI**: This dataset comes from the *Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification* dataset [19]. This dataset contains a total of 5,863 x-ray images, which are split into two classes (Normal/Pneumonia). These x-ray images were selected from pediatric patients of Guangzhou Women and Children's Medical Center, ranging from one to five years old. The filename of each image also indicates whether the patient was suffering from bacterial or viral pneumonia.
- **RNSA-PDC**: This dataset comes from the *RSNA Pneumonia Detection Challenge* [20]. This dataset contains a total of 26,684 x-ray images, which are split into two classes (Normal/Pneumonia). To gather the data required for such a large dataset, the Radiological Society of North America (RSNA) collaborated with MD.ai and the National Institutes of Health (NIH) to create a dataset of chest x-ray images which were annotated for the presence of pneumonia [21].
- **CIDC**: This dataset comes from the *COVID-19 image data collection*, an open-source dataset which is hosted on GitHub [22]. This dataset contains a total of 951 data samples, which are a mix of chest x-ray images and CT scans. Each data sample contains a number of different labels, including the findings of the radiologist, the date of the scan, and the patient.

Since the data provided in the **CIDC** dataset was a mix of chest x-ray images and CT scans, only the chest x-ray images were used in this project. The data was filtered to only include images which were labelled as a chest x-ray, then the remaining images were split into two classes (Normal/Pneumonia). The data distribution of each dataset can be seen in Table 3.1.

Dataset	Normal	Pneumonia
CXRI	1,583	4,280
RNSA-PDC	18,191	8,493
CIDC	15	509

Table 3.1: Data Distribution of Datasets

For the purpose of this project, a subset of images were selected from the RNSA-PDC dataset, which were used to create two new datasets - each containing 1000 images. This was done to create an additional client device, which would be used to test the federated learning system. It also allowed for a more balanced distribution of data between the client devices, as the majority of the data was contained in the RNSA-PDC dataset. The new datasets were named **RNSA-1** and **RNSA-2**.

Since the format of the data in each of these datasets were different, a number of pre-processing steps were required to prepare the data for training. The first step was to convert the images into a standard format, which could be used by the model. Using the format in which the **CXRI** dataset was provided, the images were organised into two folders - one for the normal images, and one for the pneumonia images.

3.2.2 Training Global Model

The global model in our federated learning framework was initialised, and trained on a proportion of the data from each device. The training was performed using a standard convolutional neural network architecture, designed to process and classify image data effectively. This training method of the global model reflects the initial stage of the federated learning process, where the global model is trained on a subset of the data from each device.

The training process utilised the TensorFlow and Keras libraries, which are widely used for machine learning and deep learning tasks. The model was trained using the Adam optimiser, which is a popular optimisation algorithm for training neural networks. The specific hyperparameters used for training the model were chosen to achieve a balance between training time and model performance, and their values can be seen in Table 3.2.

Initially, the intention was to implement a more complex model which would be able to

Table 3.2: Training Parameters for the Global Model

Parameter	**Value**
Number of Epochs	50
Batch Size	(Size of training set / Number of epochs)
Image Target Size	128×128
Classes	Binary
Optimizer	Adam
Loss Function	Binary Cross-Entropy
Metrics	Accuracy

distinguish between different types of pneumonia, such as bacterial and viral pneumonia. This however was not feasible since the information was not available for every image in the dataset. The consideration was made to remove all images which did not contain this information. After additional research, it was found that the presence of bacterial or viral pneumonia was not always visually distinguishable in chest x-ray images, and that the only way to accurately determine the type of pneumonia was to test the patient for the specific pathogen [23, 24]. In addition to this, there a high number of patients with viral pneumonia who then develop secondary bacterial infections, which can make it even more difficult to distinguish between the two types of pneumonia [25]. The model was therefore trained to classify the images as either normal or pneumonia, a binary classification task.

3.3 Identifying Non-IID Trends in Data

Identifying non-IID trends in the data is a crucial step in the development of a federated learning system. The presence of non-IID data can negatively impact federated learning models, as described in section 2.2.1. Non-IID data lead to a number of different issues, such as overfitting, poor generalisation, and slow convergence.

Non-IID data can occur due to a number of different factors. In the context of chest x-ray images, non-IID data can be caused by variations in the imaging equipment, differences in patient demographics (such as age or ethnicity), and variations in the imaging protocols which are used. These factors can lead to inconsistencies in the data, which can have a significant impact on the performance of the model. This section will explore, and highlight examples of non-IID trends which were observed in the datasets used in this project.

3.3.1 Variation in Image Quality

Inconsistencies in the image quality were observed not only between the different datasets, but also within the same dataset. This was particularly evident in the *RNSA-PDC* dataset, where inconsistencies in image brightness, contrast and noise were observed. Figure 3.1 shows examples of the variation in image quality which was observed in the dataset. Each of these images were labelled as 'Normal', however, the quality of the images varied significantly.

Figure 3.1: Variation in image quality observed in RNSA-PDC dataset, Each of these images were labelled as 'Normal'.

Further examples of this can be seen across the different datasets, where the quality of the images varied significantly. This variation in image quality can have a significant impact on the performance of the model. Figure 3.2 shows examples of the variation in image quality which were observed across the different datasets, again each of these images were labelled as 'Normal'.

Figure 3.2: Variation in image quality observed across each dataset, Each of these images were labelled as 'Normal'.

Note that the images shown in Figure 3.1 and Figure 3.2 were selected manually, and are not representative of the entire dataset. However, they are clear examples of the variation in image quality which was observed. This variation is a clear example of non-IID trends in the data, which can have a significant impact on the performance of the model.

3.3.2 Variance in Physical Features

In addition to the variation in image quality, there was a large variation in the different body types, and physical features which were present in the images. This was observed across the different datasets, where the images contained a wide range of different physical features, such as different body types, different ages and potential deformities. Examples of this can be seen in Figure 3.3.

Figure 3.3: Collection of images, collected across all datasets, with observed physical variation among patients. The patient in the left image appears to have a severe curvature of the spine. The patient in the middle image, is a very young child. The patient in the right image, is an older male, who appears to be overweight.

3.3.3 Presence of Foreign Objects

In addition to the variation in physical features, there were a number of foreign objects which were present in the images. A foreign object is defined as any object which is not part of the human body, and which is present in the image. These can appear in the form of jewellery, medical devices, or piercings.

The presence of foreign object introduces variability in the data. Each type of foreign object can have a different impact on the image, as they tend to vary significantly, in terms or shape, size and density. Foreign objects can also appear in a number of different locations in an X-ray image. For instance, necklaces and earrings can appear in the neck and head region, while piercings can appear in the chest and abdomen region, further increasing the variability in the data. Examples of foreign objects which were observed in the images can be seen in Figure 3.4.

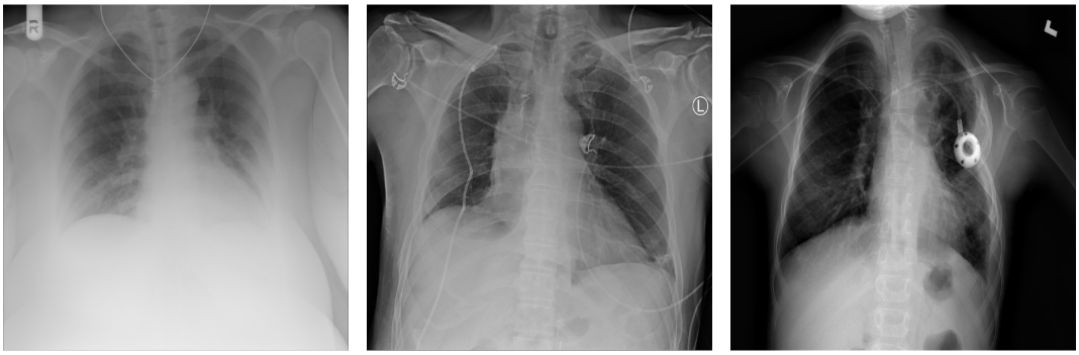

Figure 3.4: Presence of foreign objects present in the x-ray images. The patient in the left image is wearing a necklace, with a pendant in the shape of what appears to be the Barbie logo. The patient in the middle, and right images both have medical devices attached to their bodies.

Due to how uncommon and seemingly random the presence of foreign objects are in the data, their specific impact on model performance can be difficult to quantify, and even more difficult to mitigate. In a hypothetical situation where everyone who is wearing a visually distinct necklace has pneumonia, the model may learn to associate the presence of the necklace with the presence of pneumonia. This would be an example of a non-IID trend in the data, which could have a significant impact on the performance of the model.

3.3.4 Inconsistencies in X-Ray Image Annotation

Image annotation is a common practice in the field of Radiography, in which labels or annotations are added to the image to improve understanding, add context, or to highlight specific areas of interest. Generally speaking, annotations are mainly used to label the left, or right side of the body, so that people viewing the image can easily identify which side of the body they are looking at.

In the datasets which were used in this project, there were a number of inconsistencies in the image annotations. This was particularly evident in the *CIDC* dataset, where in some cases, additional arrows and text were added to highlight the presence of pneumonia. Figure 3.5 shows examples of these annotations. The addition of these arrows and text can be useful for the human eye to help locate the area of interest at a quick glance, especially for people who are not medical personnel. However, they can negatively impact the performance of the model, as the model may learn to associate the presence of these annotations with the presence of pneumonia, which may not always be the case.

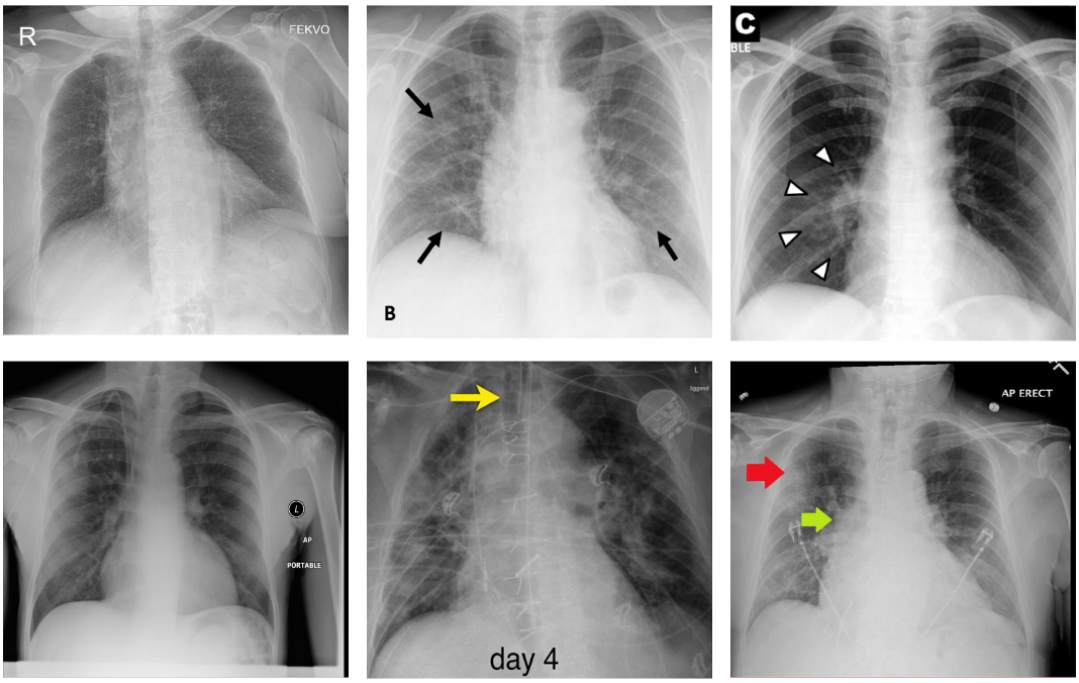

Figure 3.5: Evidence of inconsistencies in the image annotations, which were observed in the CIDC dataset. The images on the top left and bottom left, are examples of non-intrusive annotations. Where each annotation is present to denote the right, and left side of the body respectively. The remaining images are clear examples of intrusive annotations, which contain the addition of text and arrows to highlight the presence of pneumonia.

3.4 Creation of Personalised Models

3.4.1 Personalisation Strategy

In an ideal world, after the initial federated learning process has been completed (outlined in section 3.2.2), the global model would be able to make accurate predictions on the local data from each device. However, the non-IID nature of the data (as highlighted in section 3.3) can have a significant, negative impact on the performance of the model.

To counteract this, an additional personalisation layer was added to each local device, which occurs after the initial training of the global model. This personalisation layer is aimed to produce a personalised model for each local device, which are fine-tuned to their local datasets. Given the state-of-the-art personalisation techniques which have been discussed in section TODO, the personalisation layer operates by following the steps below:

1. **Global Model Distribution**: The global model which has been produced by the initial federated learning process is distributed to each local device.

2. **Image Augmentation**: A number of image augmentation techniques are applied to the local data, which are used to reduce the impact of any outliers in the data.

3. **Synthetic Data Generation**: Synthetic data is generated based on the local data, which is designed to increase the size of the training set. This is done to prevent overfitting, which can happen when the size of the training set is too small.

4. **Duplicate CNN**: A duplicate of the convolutional neural network (CNN) which was used to train the global model is created.

5. **Transfer Learning**: A new model is created using the duplicate CNN, which is then initialised with the weights of the global model. This model is then trained on the local data, along with the synthetic data which was generated in the previous step.

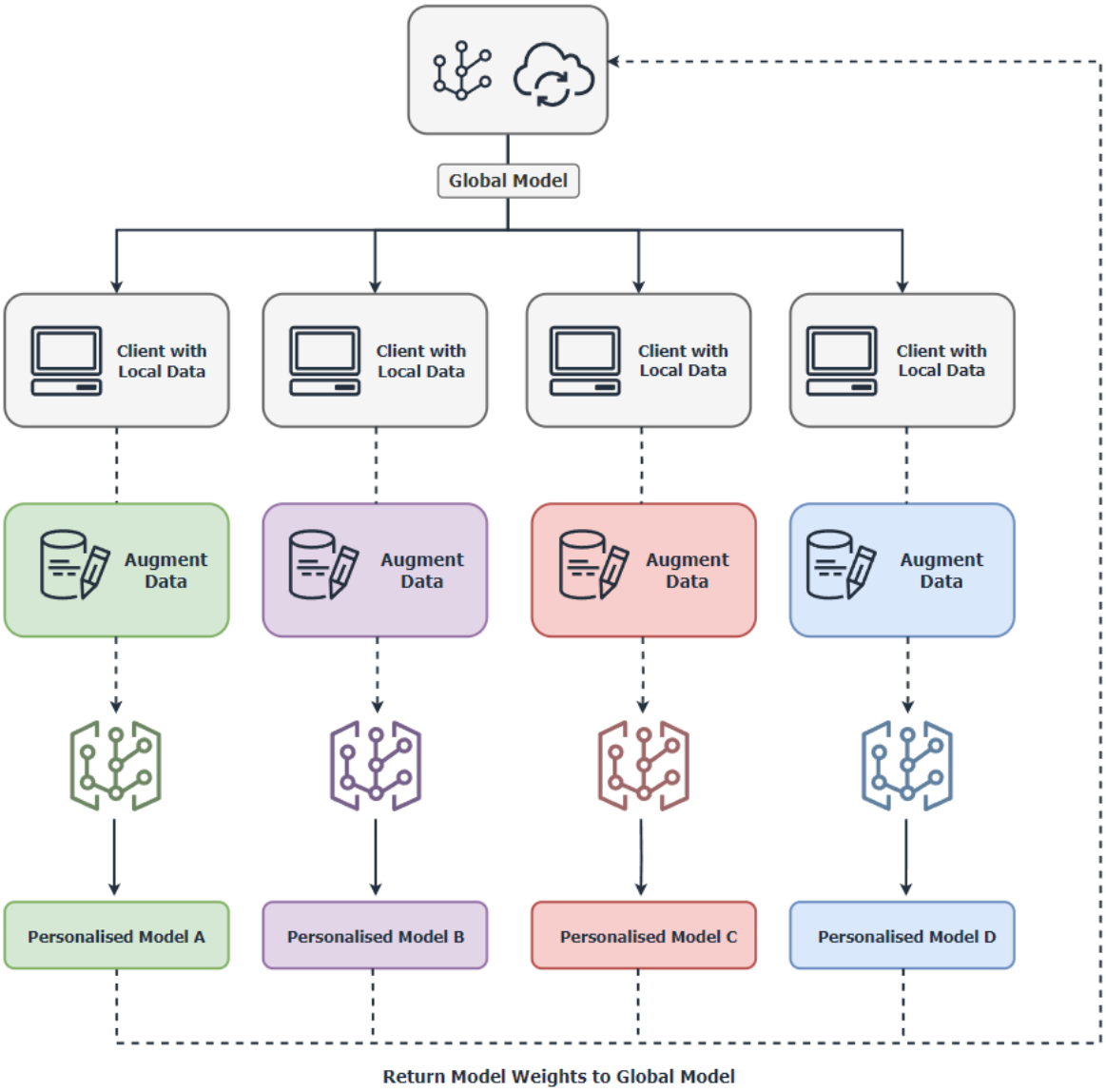

Figure 3.6: Personalisation Strategy

3.4.2 Image Augmentation

Image Augmentation was performed in order to reduce the impact of any outliers in the data. This was to ensure that the model was able to learn the underlying patterns in the data, which could be negatively impacted by the presence of outliers.

The first step in the image augmentation process was to resize the images to a standard size, which was 128×128 pixels. This was done to ensure that all of the images were the same size, which is a requirement for training the model.

Next, the images were converted to greyscale. This was a necessary step, as there were some outliers in the datasets in which the images were not in greyscale. By converting the images to greyscale, we were able to ensure that the model was only learning from the intensity of the pixels, rather than the colour of the images.

A number of additional image augmentation techniques were also applied in the creation of synthetic data, which is discussed in section 3.4.3.

3.4.3 Synthetic Data Generation

Synthetic data is generated based on the local data. This is done to increase the size of the training set with new data samples, which are designed to prevent overfitting. Overfitting can occur when the size of the training set is too small, and the model learns from the noise in the data, rather than the underlying patterns.

The synthetic data is generated from a subset of the local data, where each image has a 50% chance of being selected. An image augmentation pipeline is then employed, which applies a number of different image augmentation techniques to the selected images. These steps are designed to simulate the effects of different imaging conditions.

The specific image augmentation techniques which were used in the pipeline are shown in Table 3.3. The design of this image augmentation pipeline was heavily by the work of Schaudt et al. [26], who used a similar approach to increase image variation, and to reduce overfitting while training their model.

Technique	**Description**	**Probability**
Rotation	Perform a random rotation on the image, between -15 and 15 degrees.	1.0
Scale	Zoom in/out on the image, from a random value between -10% and 10%.	1.0
Horizontal Skew	Skew the image horizontally, with a random skew factor between -0.15 and 0.15.	0.5
Vertical Skew	Skew the image vertically, with a random skew factor between -0.15 and 0.15.	0.5
Brightness/ Contrast	Adjust the brightness of the image, with a random factor between 0.8 and 1.2. Then, adjust the contrast of the image by a random factor between 0.8 and 1.2	0.9
Sharpen/ Blur	Randomly apply one of the following adjustments: - Gaussian Blur: Apply a subtle gaussian blur to the image, with a random radius between 0.1 and 1. - Sharpness: Adjust the sharpness of the image, with a random value factor between 1.0 and 1.2.	0.9

Table 3.3: Proposed Image Augmentation Pipeline, which is used to generate synthetic data for the personalisation layer.

3.4.4 Transfer Learning

Using an identical convolutional neural network (CNN) to the one which was used to train the global model, a new model was created. This model was then initialised with the weights of the global model. This process is known as transfer learning, as described in section 2.3.2. The model was then trained on the local data, along with any synthetic data which was generated in the previous step.

The concept behind transfer learning is that the model is able to learn the underlying patterns in the data more effectively, since it had already been trained on a similar dataset. This allows the model to converge more efficiently, and to improve the accuracy of the predictions which are made on the local data.

3.4.5 Personalised Model Training

Todo

3.5 Testing

In order to evaluate the performance of the models, each dataset needed to be split into a training, validation, and test set. For the **CXRI** dataset, the data was already split into a training and test set, so no further action was required. For the remaining datasets, 70% of the data was used for training, 10% was used for validation, and 20% was used for testing.

The purpose of the training set is to train the global model to fit the data which is present on each device. The synthetic data was generated from the training set, then used to add additional data samples to the training set.

The validation set was used to evaluate the performance of the model during training, with the main goal of preventing overfitting.

The test set was used to evaluate the performance of the model after training had been completed, by testing the model on completely unseen data. The performance of the model on the test set was used to indicate how well the model would perform in a real-world scenario, where the data is not known in advance.

Only data from the original datasets were used for testing and validation. No synthetic data was included in either of these sets, as the purpose of the synthetic data was to increase the

size of the training set. Running the model on synthetic data would not provide an accurate representation of how the model would perform on real-world data, as the synthetic data was generated from the training set (which the model had already seen).

An accuracy score was calculated at the end of the training process, which was used to evaluate the performance of the model. The equation for calculating the accuracy score is shown in Equation 3.1.

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{False Positives} + \text{True Positives} + \text{True Negatives} + \text{False Negatives}} \tag{3.1}$$

The main goal of this experiment is to evaluate the performance of the models, and to determine whether the personalisation layer was effective in improving their performance. To do this, training was performed on the global model, and then on the personalised models, with varying rounds of training, and varying amounts of synthetic data. This was done to determine the optimal number of rounds of training, and the optimal amount of synthetic data which should be used to train the model. The specific number of training rounds and synthetic data which were used in the experiment can be seen in Table 3.4.

Epochs	2	5	10
Synthetic Data	No Synthetic Data	50% of Original Data Size	100% of Original Data Size

Table 3.4: Range of Values for Model Training, used to determine the optimal number of training rounds, and the optimal amount of synthetic data which should generated, and added to the training set.

3.6 Results & Analysis

After training the global model in the initial federated learning process, the model was then tested on each of the datasets. The results of this testing can be seen in Table 3.5.

Dataset	Accuracy
CXRI	78.4%
CIDC	55.25%
RNSA-1	65.5%
RNSA-2	66.2%
Average Score	66.09%

Table 3.5: Accuracy Scores produced by the Global Model on each dataset.

After the global model had been trained, the personalisation layer was added to each local device. The model was then trained in accordance to the proposed personalisation strategy, using varying amounts of synthetic data and number of epochs.

Additional Synthetic Data	2 Epochs	5 Epochs	8 Epochs
No Synthetic	75%	80%	85%
50% of Original Data	82%	87%	88%
100% of Original Data	84%	88%	89%

Table 3.6: Personalised Model Accuracy Scores on CXRI Dataset

Additional Synthetic Data	2 Epochs	5 Epochs	8 Epochs
No Synthetic	53%	58%	62%
50% of Original Data	60%	64%	66%
100% of Original Data	65%	68%	70%

Table 3.7: Personalised Model Accuracy Scores on CIDC Dataset

Additional Synthetic Data	2 Epochs	5 Epochs	8 Epochs
No Synthetic	65%	70%	75%
50% of Original Data	72%	76%	78%
100% of Original Data	74%	77%	79%

Table 3.8: Personalised Model Accuracy Scores on RNSA-1 Dataset

Additional Synthetic Data	2 Epochs	5 Epochs	8 Epochs
No Synthetic	66%	71%	76%
50% of Original Data	73%	77%	79%
100% of Original Data	75%	78%	80%

Table 3.9: Personalised Model Accuracy Scores on RNSA-2 Dataset

Additional Synthetic Data	2 Epochs	5 Epochs	8 Epochs
No Synthetic	64.75%	69.75%	74.5%
50% of Original Data	71.75%	76.0%	77.75%
100% of Original Data	74.5%	77.75%	79.5%

Table 3.10: Average Personal Model Accuracy Scores across all Datasets

4 Evaluation & Critical Analysis

4.1 Analysis of Results

When using the optimal hyperparameters for each model (which were obtained in section 3.6), the personalised model outperforms the global model for every data set. The personalised models on average, outperformed the global model by 13.16%.

The largest improvement was seen with the CIDC dataset, where the personalised model outperformed the global model by 14.75%. The smallest improvement was seen with the CXRI dataset, where the personalised model outperformed the global model by 10.6%.

The differences in these improvements highlight the influence of the data size and quality on the performance of the models. The CIDC dataset, contained the least amount of data, while also containing the most examples of non-IID trends. The dataset was an aggregation of open-source data, meaning that the data had more inconsistencies and outliers. This is reflected in the fact that the global model had the lowest accuracy on the CIDC dataset when compared to the other datasets, which can be seen in Table 3.5. The personalised model was able to outperform the global model by a larger margin, as the personalised model was able to adapt to the non-IID trends in the data.

Alternatively, the CXRI dataset, contained the most data samples, with the least amount of variation. This dataset contained chest x-rays images which were selected from pediatric patients of Guangzhou Women and Children's Medical Center, ranging from 1-5 years of age [19]. The data was all collected from a single source, likely using the same equipment and procedures. While there was still evidence of non-IID trends in the data, the data was more consistent when compared to other datasets, which were aggregated from multiple sources. This also explains the higher accuracy of the global model on the CXRI dataset when compared to the other datasets, which can be seen in Table 3.5. The personalised model was still able to outperform the global model, but by a smaller margin, as the global model was already performing well on the CXRI dataset.

4.2 Evaluation of Methodology

4.2.1 Machine Learning Techniques

This project explores the implementation of federated learning for detecting the presence of pneumonia in chest X-ray images. The use of federated learning allows for the models to be trained without the need for centralised data storage, allowing them to adhere to privacy concerns and regulations such as GDPR [2].

This project utilised convolutional neural networks (CNNs) for the classification of the chest X-ray images. CNNs are a popular choice for image classification due to their efficiency in handling image recognition tasks. The use of CNNs in this project allowed for the models to extract features from chest X-ray images, these features were then used to classify the images as either normal or pneumonia.

The use of techniques like transfer learning and synthetic data generation allowed for the model training to be enhanced, without compromising the privacy of the data. In particular, transfer learning was used to fine-tune the models locally, using the global model as a starting point. This allowed for the models to be trained with less data, while still achieving high accuracy. Synthetic data generation was used to increase the amount of data available for training, which was particularly useful for the CIDC dataset, which contained the least amount of data. These techniques were crucial in ensuring that the personalised models were able to counteract the effects of non-IID data.

4.2.2 Identification of Non-IID Data

The main goal of this project was to investigate the effects of non-IID data on the performance of federated learning models, particularly in the context of medical image classification.

A number of different trends were identified in the data, which could be classified as non-IID. These trends included:

- Variation in image quality.
- Variation in patient demographics.
- Presence of foreign objects in images.

- Inconsistencies in X-Ray annotation.

While these trends were identified in the data, with examples provided in section 3.3, it is not clear the extent to which these observed trends impacted the performance of the models.

There was clearly presence of non-IID data in the datasets, as the global model struggled to achieve high accuracy on the local datasets (see section 3.6). The personalised models were able to outperform the global model, which suggests that the personalised models were able to adapt to the non-IID trends in the data. However, it is not clear how each of these trends impacted the performance of the models. It would have been beneficial to conduct a more detailed analysis of the impact of each of these trends on the performance of the models, by quantifying the divergence caused by each trend using techniques such as Jensen-Shannon divergence [27].

4.2.3 Personalisation Strategies

The main techniques which were used to personalise the models were synthetic data generation and transfer learning. Transfer learning was used to fine-tune the models locally, using the global model as a starting point. This allowed for the models to be trained with less data, while still achieving high accuracy. Synthetic data generation was used to increase the amount of training data, which proved to be particularly useful for the CIDC dataset, which contained the least amount of data.

The goal of these techniques were to reduce the effect of non-IID data on the performance of the models. The results show that the personalised models were able to outperform the global model, which suggests that these techniques were successful in achieving this goal. However, not every non-IID trend, which were discussed in section 3.3, were accounted for in the personalisation strategies. For example, the presence of foreign objects in images was not accounted for at any point, it would have been beneficial to not only quantify the effect of these trends on the performance of the models, but also to investigate how these trends could be accounted for in the personalisation strategies. One possible approach to this would be to run a 2-step classification process, as described by Gomes et al. [14], where the first step would be to identify features in the image, such as body parts or foreign objects. The second step would then involve using these features as an additional input to detect the presence of pneumonia.

In terms of the synthetic data generation, an image augmentation pipeline was used to generate synthetic data. This pipeline was heavily inspired by the work of Schaudt et al. [28], in which an image augmentation pipeline was employed to increase image variation,

and to reduce overfitting that could occur when training models on small datasets.

Other techniques for the synthetic data generation could have been explored. For example, in the same paper by Schaudt et al. [28], they also explored the use of generative adversarial networks (GANs), along with diffusion models, to generate synthetic data. The GAN model consists of two components, a generator and a discriminator. The generator generates synthetic data, while the discriminator tries to distinguish between real and synthetic data [29]. The training process alternates between optimising the generator, to generate more realistic data, and the discriminator, to better distinguish between real and synthetic data. This training process continues until the generator is capable of generating data which is indistinguishable from real data. This technique could have been explored in this project, but was not due to time constraints.

Other techniques for personalisation could have been explored. For example, the use of multi-task learning [13], which allows for the models to learn multiple tasks simultaneously, could have been explored. This technique could have proven useful in identifying specific non-IID trends in the data, such as the presence of foreign objects in images, or inconsistencies in X-Ray annotation.

The personalisation strategy which was proposed in this project, the personalised models achieved the highest accuracy after 5-10 epochs, and it is assumed that overfitting does not occur. The addition of synthetic data generation was used to reduce the risk of overfitting [30], by increasing the amount of training data. Presumably, if the models were to be trained for more epochs, overfitting would occur. This could have been explored further, by training the models for more epochs. This would have allowed for the investigation of the effects of overfitting, and how effective the synthetic data generation was in reducing the risk of overfitting.

5 Conclusion

5.1 Overview

This dissertation is aimed to investigate the application of personalised federated learning for the detection of pneumonia in chest X-ray images. The primary goal of this project was to explore the challenges posed by non-IID data across different distributed datasets.

The negative impact of non-IID data on the performance of federated learning models was explored, and proven to have a detrimental effect on the performance of the global model. This is due to the global model being unable to adapt to the non-IID trends in the data, which results in the model weights diverging from the global model.

These negative effects are particularly prevalent in the context of medical data. Medical data is highly prone to non-IID trends, due to the variation in patient demographics, image quality, and data collection procedures. The presence of non-IID data in medical data can lead to models making incorrect predictions, which can have serious consequences in a clinical setting.

Strategies to mitigate the effects of non-IID data were explored, with a focus on personalisation techniques. Personalisation strategies such as transfer learning and synthetic data generation were used in the final implementation to fine-tune the models locally, and to increase the amount of training data. These techniques were successful in improving the performance of the models, with the personalised models outperforming the global model across all datasets, with an average improvement of 13.16%.

5.2 Future Work

While the proposed method was successful in improving the performance of the models, there are still areas which could be explored further. One area which could be explored is the

use of more advanced personalisation techniques, such as multi-task, to adapt models to specific non-IID trends in the data without compromising the learning of other trends.

Another area which could be explored is the use of more advanced synthetic data generation techniques, such as generative adversarial networks (GANs). GANs could be used to generate more realistic and unique synthetic data, which could be used to further enhance the performance of the models. This was not explored in this project due to time constraints, but sounds like a promising avenue for future work.

Another challenge which is specific to federated learning which was not explored in this project is the issue of transmission/computation costs. In a real-world scenario, the transmission of model updates between clients and the server can be costly, particularly in the case of large models. Computational costs can also be high, particularly for clients with limited computational resources. Investigating the impact of these costs on the performance of the models, and exploring techniques to reduce these costs, is definitely an area where future work could be focused.

The impact of overfitting on the models could also be explored further. The models were trained for a small number of epochs to prevent overfitting, but it is not clear how the models would perform if trained for more epochs. Investigating the effects of overfitting, and how effective the synthetic data generation was in reducing the risk of overfitting, could be an interesting area for future work.

Finally, the impact of non-IID data on the performance of the models could be explored in more detail. While the presence of non-IID data was identified in the datasets, it is not clear how each of these trends specifically impacted the performance of the models. Quantifying the divergence caused by non-IID using techniques such as Jensen-Shannon divergence could provide more insight into the effects of non-IID data on the models.

5.3 Reflection

Overall, this research journey has been both challenging and rewarding, providing me with a great learning experience. I have gained a deeper understanding of federated learning, and the unique challenges posed by non-IID data. I have also developed my technical skills, particularly in the areas of machine learning and data analysis.

One of the main challenges I faced during this project was in terms of data collection. At the beginning stages of the project, I struggled to find what I considered suitable datasets for the task at hand. This led to a delay in the project timeline, as I had to spend more time

searching for and cleaning the data. This was quite honestly a huge waste of time, as I was searching for the "perfect" dataset, which does not exist. In hindsight, the imperfect datasets which I eventually settled on were more than sufficient for the task at hand. This was a valuable lesson in the importance of being adaptable and flexible in research, and not getting too caught up in the details.

Another challenge I had faced was choosing the right model architecture for the task at hand. I initially wanted to use a more complex model architecture, and fell down the rabbit hole of researching the state-of-the-art model architectures for image classification. This led to a lot of confusion and wasted time, as I was trying to implement models which were far too complex for the task at hand. In the end, I settled on a simple convolutional neural network, which was more than sufficient for the task at hand. This was another valuable lesson in the importance of simplicity in research, and not getting too caught up in the details.

While there are many things I would do differently if I were to start this project from scratch, I am proud of the work I have produced. Overall, the results from this project were promising - Highlighting the prevalence of non-IID data in medical contexts, and proving that personalisation techniques can be effective in mitigating the effects of non-IID data.

Bibliography

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, p. 84–90, may 2017. [Online]. Available: https://doi-org.elib.tcd.ie/10.1145/3065386

[2] P. Voigt and A. v. d. Bussche, *The EU General Data Protection Regulation (GDPR): A Practical Guide*, 1st ed. Springer Publishing Company, Incorporated, 2017.

[3] Centers for Medicare & Medicaid Services, "The Health Insurance Portability and Accountability Act of 1996 (HIPAA)," Online at http://www.cms.hhs.gov/hipaa/, 1996.

[4] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, A. Singh and J. Zhu, Eds., vol. 54. PMLR, 20–22 Apr 2017, pp. 1273–1282. [Online]. Available: https://proceedings.mlr.press/v54/mcmahan17a.html

[5] D. van Esch, E. Sarbar, T. Lucassen, J. O'Brien, T. Breiner, M. Prasad, E. Crew, C. Nguyen, and F. Beaufays, "Writing across the world's languages: Deep internationalization for gboard, the google keyboard," *CoRR*, vol. abs/1912.01218, 2019. [Online]. Available: http://arxiv.org/abs/1912.01218

[6] H. Zhu, J. Xu, S. Liu, and Y. Jin, "Federated learning on non-iid data: A survey," *CoRR*, vol. abs/2106.06843, 2021. [Online]. Available: https://arxiv.org/abs/2106.06843

[7] C. Health, "Aging and chronic disease - the american tsunami of bad health," Oct 2019. [Online]. Available: https://www.prnewswire.com/news-releases/aging-and-chronic-disease--the-american-tsunami-of-bad-health-300939034.html

[8] Y. Huang, L. Chu, Z. Zhou, L. Wang, J. Liu, J. Pei, and Y. Zhang, "Personalized federated learning: An attentive collaboration approach," *CoRR*, vol. abs/2007.03797, 2020. [Online]. Available: https://arxiv.org/abs/2007.03797

[9] H. Zhu, J. Xu, S. Liu, and Y. Jin, "Federated learning on non-iid data: A survey," *CoRR*, vol. abs/2106.06843, 2021. [Online]. Available: https://arxiv.org/abs/2106.06843

[10] L. Su, J. Xu, and P. Yang, "A non-parametric view of fedavg and fedprox: beyond stationary points," *J. Mach. Learn. Res.*, vol. 24, no. 1, mar 2024.

[11] S. P. Karimireddy, S. Kale, M. Mohri, S. J. Reddi, S. U. Stich, and A. T. Suresh, "Scaffold: Stochastic controlled averaging for federated learning," 2021.

[12] K. Woodward, E. Kanjo, D. J. Brown, and T. M. McGinnity, "On-device transfer learning for personalising psychological stress modelling using a convolutional neural network," *CoRR*, vol. abs/2004.01603, 2020. [Online]. Available: https://arxiv.org/abs/2004.01603

[13] M. Crawshaw, "Multi-task learning with deep neural networks: A survey," *CoRR*, vol. abs/2009.09796, 2020. [Online]. Available: https://arxiv.org/abs/2009.09796

[14] D. Gomes and I. A. Lawal, "X-ray image classification using two-step densenet classifiers," in *Proceedings of the 14th PErvasive Technologies Related to Assistive Environments Conference*, ser. PETRA '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 550–555. [Online]. Available: https://doi-org.elib.tcd.ie/10.1145/3453892.3461632

[15] G. Huang, Z. Liu, and K. Q. Weinberger, "Densely connected convolutional networks," *CoRR*, vol. abs/1608.06993, 2016. [Online]. Available: http://arxiv.org/abs/1608.06993

[16] Q. Wu, X. Chen, Z. Zhou, and J. Zhang, "Fedhome: Cloud-edge based personalized federated learning for in-home health monitoring," *CoRR*, vol. abs/2012.07450, 2020. [Online]. Available: https://arxiv.org/abs/2012.07450

[17] O. V. Shcherbakov, I. N. Zhdanov, and Y. A. Lushin, "A convolutional autoencoder as a generative model of images for problems of distinguishing attributes and restoring images in missing regions," *J. Opt. Technol.*, vol. 82, no. 8, pp. 528–532, Aug 2015. [Online]. Available: https://opg.optica.org/jot/abstract.cfm?URI=jot-82-8-528

[18] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, p. 321–357, Jun. 2002. [Online]. Available: http://dx.doi.org/10.1613/jair.953

[19] D. S. Kermany, K. Zhang, and M. H. Goldbaum, "Labeled optical coherence tomography (oct) and chest x-ray images for classification," 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:126183849

[20] A. Stein, M. C. Wu, C. Carr, G. Shih, J. Dulkowski, kalpathy, L. Chen, L. Prevedello, M. Kohli, MD, M. McDonald, Peter, P. Culliton, S. H. MD, and T. Xia, "Rsna pneumonia detection challenge," 2018. [Online]. Available: https://kaggle.com/competitions/rsna-pneumonia-detection-challenge

[21] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," *CoRR*, vol. abs/1705.02315, 2017. [Online]. Available: http://arxiv.org/abs/1705.02315

[22] J. P. Cohen, P. Morrison, L. Dao, K. Roth, T. Q. Duong, and M. Ghassemi, "Covid-19 image data collection: Prospective predictions are the future," *arXiv 2006.11988*, 2020. [Online]. Available: https://github.com/ieee8023/covid-chestxray-dataset

[23] K. Stefanidis, E. Konstantelou, G. T. Yusuf, A. Oikonomou, K. Tavernaraki, D. Karakitsos, S. Loukides, and I. Vlahos, "Radiological, epidemiological and clinical patterns of pulmonary viral infections," *European Journal of Radiology*, vol. 136, p. 109548, Mar. 2021. [Online]. Available: http://dx.doi.org/10.1016/j.ejrad.2021.109548

[24] D. M. Musher, I. L. Roig, G. Cazares, C. E. Stager, N. Logan, and H. Safar, "Can an etiologic agent be identified in adults who are hospitalized for community-acquired pneumonia: Results of a one-year study," *Journal of Infection*, vol. 67, no. 1, p. 11–18, Jul. 2013. [Online]. Available: http://dx.doi.org/10.1016/j.jinf.2013.03.003

[25] P. Manohar, B. Loh, R. Nachimuthu, X. Hua, S. C. Welburn, and S. Leptihn, "Secondary bacterial infections in patients with viral pneumonia," *Frontiers in Medicine*, vol. 7, 2020. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fmed.2020.00420

[26] D. Schaudt, C. Späte, R. Schwerin, M. Reichert, M. Schwerin, M. Beer, and C. Kloth, "A critical assessment of generative models for synthetic data augmentation on limited pneumonia x-ray data," *Bioengineering*, vol. 10, p. 1421, 12 2023.

[27] J. Deasy, N. Simidjievski, and P. Liò, "Constraining variational inference with geometric jensen-shannon divergence," *CoRR*, vol. abs/2006.10599, 2020. [Online]. Available: https://arxiv.org/abs/2006.10599

[28] D. Schaudt, C. Späte, R. von Schwerin, M. Reichert, M. von Schwerin, M. Beer, and C. Kloth, "A critical assessment of generative models for synthetic data augmentation

on limited pneumonia x-ray data," *Bioengineering*, vol. 10, no. 12, 2023. [Online]. Available: https://www.mdpi.com/2306-5354/10/12/1421

[29] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, p. 139–144, oct 2020. [Online]. Available: https://doi-org.elib.tcd.ie/10.1145/3422622

[30] J. A. Solworth, "Epochs," *ACM Trans. Program. Lang. Syst.*, vol. 14, no. 1, p. 28–53, jan 1992. [Online]. Available: https://doi-org.elib.tcd.ie/10.1145/111186.116785

A1 Appendix

Source code for the project can be found on GitHub:
`https://github.com/tomrobb/Personalised-Federated-Learning-Pneumonia`