Federated Learning in the Detection of COVID-19 in Patient CT-Scans: A Practical Evaluation of External Generalisation

School of Computer Science & Applied Mathematics
University of the Witwatersrand

Korstiaan Wapenaar (1492459)

Supervised by Dr Pravesh Ranchod

August 23, 2023

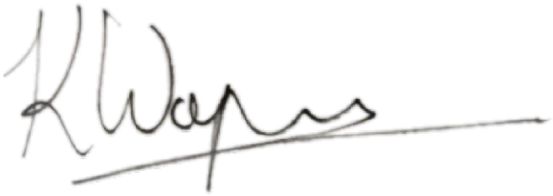
A research report submitted to the Faculty of Science, University of the Witwatersrand, Johannesburg, in partial fulfilment of the requirements for the degree of Master of Science

Abstract

This research explores the practical utility of using convolutional neural networks in a federated learning architecture for COVID-19 diagnostics using chest CT-scans, and whether federated learning models can generalise to data from healthcare facilities that did not participate in training. A model that can generalise to these healthcare facilities could provide lower-resourced or over-utilised facilities with access to supplementary diagnostic services. Eleven models are trained using a modified VGG-16. The models are trained using data from five 'sites': four sites are single healthcare facilities and the fifth site is a composite of data from a variety of healthcare facilities. Eleven models are trained, evaluated and compared: five 'independent models' are each trained with data from a single site; three 'global models' are trained using centrally pooled data from a variety of sites; three 'federated models' are trained using a federated averaging approach. The site with composite data is held-out and never included in training the federated and global models. With the exception of this composite site, all models achieve a test accuracy of at least 0.93 when evaluated using test data from the sites used in training these models. All models are then evaluated using data from the composite site. The global and federated models achieve a 0.5 to 0.6 accuracy for the composite site, indicating that the model and training regime is unable to achieve useful accuracies for sites non-participant in training. The federated models are therefore not accurate enough to motivate a healthcare facility decision maker to use the federated models as an alternative or supplementary diagnostic tool to radiographers, or to developing their own independent model. Evaluation of the results suggests that high-quality and consistent image pre-processing may be a necessary precondition for the task.

Declaration

I, Korstiaan Wapenaar, hereby declare the contents of this research report to be my own work. This report is submitted for the degree of Master of Science at the University of the Witwatersrand. This work has not been submitted to any other university, or for any other degree.



Contents

Preface
Abstract i
Declaration ii
Table of Contents iii
List of Figures v
List of Tables vi

1 Introduction **1**
1.1 Introduction 2
1.2 Research Objectives 3
1.3 Background 4
1.3.1 COVID-19 and AI 4
1.3.2 Data Protection and Healthcare 5
1.3.3 Federated Learning, Transfer Learning and Data Scarcity 6

2 Literature Review **8**
2.1 Convolutional Neural Networks and VGG16 8
2.2 Federated Learning 13
2.3 Federated Learning and CNN for Diagnostics 15
2.4 Summary of Existing Research 16

3 Data and Methodology **18**
3.1 Data 18
3.2 Methodology 21
3.2.1 Data Preparation 21
3.2.2 Model 22
3.2.3 Testing Regime and Hyperparameters 23

4 Results **28**
4.1 Independent Models 28
4.1.1 Independent Model: Site 10_2 30
4.1.2 Independent Model: Site 10_1 30
4.1.3 Independent Model: Site 7 31
4.1.4 Independent Model: Site 2 31
4.1.5 Independent Model: Site 1 31
4.2 Global Models 31

4.2.1 Global Model 1: Site 10_1 and 10_2 33
4.2.2 Global Model 2: Site 10_1 & Site 10_2 & Site 7 33
4.2.3 Global Model 3: Site 10_1 & Site 10_2 & Site 7 & Site 1 33
4.3 Federated Models 34
4.3.1 Federated Model 1: Site 10_1 and 10_2 36
4.3.2 Federated Model 2: Site 10_1 & Site 10_2 & Site 7 36
4.3.3 Federated Model 3: Site 10_1 & Site 10_2 & Site 7 & Site 1 36

5 Discussion **37**
5.1 Results 37
5.2 Further Research 39

6 Conclusion **41**

A Supplementary Model Details **42**

References **53**

List of Figures

2.1 The VGG16 Architecture [Ferguson *et al.* 2018] 9
2.2 The VGG family of CNN [Tsang 2014] 10
2.3 A Stylised Federated Learning Network [Yang *et al.* [2019a]] 13
2.4 A Federated Learning Approach to COVID-19 Diagnostics [Yang *et al.* 2021b] 15

3.1 Lesion annotations of chest CT scans for COVID-19 positive patients . . . 19
3.2 Sample of Site Images Used in Training 20
3.3 Sample of Site 2 Images 21
3.4 Sample of Images Before and After Processing 22

4.1 Independent Models Training and Loss 30
4.2 Global Models Accuracy and Loss 33
4.3 Federated Learning Models Accuracy and Loss 35

5.1 Global Model 3 Training and Validation Accuracy 39

List of Tables

3.1 Source Data Overview 19
3.2 Input Data Used in Training 20
3.3 Customised VGG16 Network Architecture 23
3.4 Model Architecture Evaluation 24
3.5 Training Regime Overview 26

4.1 Independent Model Test Accuracy Matrix 29
4.2 Independent Model Average External Test Accuracy 29
4.3 Average Test Accuracy per Site for Models Trained on Other Site Data. . 30
4.4 Global Model Internal Test Accuracy 31
4.5 Global Model Site Specific Internal and External Test Accuracy 32
4.6 Global Model Average External Test Accuracy 32
4.7 Federated Model Internal Test Accuracy 34
4.8 Federated Model Site Specific Internal and External Test Accuracy 34
4.9 Federated Model Average External Test Accuracy 35

Chapter 1

Introduction

This research report illustrates the opportunities and limits of using federated learning to detect COVID-19 in chest CT-Scans. Federated learning is a privacy preserving technology that can use CT scans from a variety of healthcare facilities in training. This research explores whether a model trained using a federated architecture can generalise to scans from sites that did not participate in training. A model that can generalise to scans from facilities that did not participate in training would provide under-resourced or over-utilised healthcare facilities with an alternative diagnostic tool to alleviate pressure on and increase throughput of facility radiologists. A federated model would require superior levels of accuracy to radiologists or to an independent model developed only using the facility's own data to be practically and safely useful.

This research report is divided into 6 chapters. The first chapter provides an introduction to the research. The second chapter reviews and summarises relevant literature. This includes convolutional-neural networks (CNN), federated learning (FL), the combination of CNN and FL, and their application in healthcare diagnostics. The third chapter details the data and methodology. It also details how the model's results are evaluated. The fourth chapter presents the results of the research. These results are discussed in chapter six which also identifies areas of further research. This is followed by a brief conclusion. The appendix is divided into two sections: a tabularised summary of the literature review, and; additional details regarding CNN and FL which supplements the research.

This chapter provides an introduction to the research report. It defines the objectives of the research and how this contributes to the literature. This chapter then provides background to the research by detailing the implications of COVID-19 on individuals and society, and how artificial intelligence was used to help mitigate these individual and social implications. This section explores how disruptive technologies such as federated learning enable COVID-19 diagnostics in an environment of scarce data while ensuring compliance with data sharing obligations required by data protection legislation.

1.1 Introduction

Privacy preserving technologies are important mechanisms for developing diagnostic solutions for healthcare while maintaining the privacy of the underlying data subjects. Medical data is often legally classified as 'special personal data'. This classification limits how freely medical data can be shared, impacts the extent to which medical data is available for research, and creates obligations for its use in research. The subsequent data scarcity impedes research in the field.

Federated learning is a privacy preserving machine learning technique that produces an openly-accessible, shared model that is trained from a set of local models. Each local model is trained on an independent and segregated dataset. After each epoch, the updates of the local models are averaged to update the parameters in the shared model until convergence is achieved. This federated averaging approach allows for the development of a diagnostic solution that leverages multiple sources of data without requiring data sharing. Convolutional neural networks (CNN) can be used within a federated learning framework. CNNs are a means of analyzing and classifying an image. CNNs can therefore be used to diagnose COVID-19 from CT scans.

There is a small but growing base of research into the use of federated learning for the detection of COVID-19 in chest CT scans. This research uses a novel dataset of COVID-19 CT scans that are segregated by site to evaluate the feasibility of using federated learning for the diagnosis of COVID-19 by sites that did not participate in training.[1]. This research employs a comprehensive regime of experimentation by comparing the accuracy of three sets of models: independent models which are each trained using data from a single site; global models which are trained using data centrally pooled from multiple sites; federated models which are trained using a federated averaging approach. Three global models and three federated models are evaluated. Each introduces an additional site to training to evaluate the impact of including more sites in training. Each global model has a federated counterpart that has been trained with the same sites. This allows for investigation into the implications of a federated learning approach, and whether or not it can achieve sufficient levels of accuracy to warrant a healthcare facility using the model as a complement to radiographers, or an alternative to developing their own model.

All models are evaluated in terms of internal and external test accuracies. Internal test accuracy refers to a model's accuracy when evaluated on test data from the site(s) used in training. External test accuracy refers to a model's accuracy when evaluated on test data from sites not used in training. External test accuracy indicates the ability for the model to generalise to new sites. This has important public health implications as it would enable a wide variety of sites to make use of the federated model.

[1]Data used in the preparation of this article were obtained from the COVID-19 DATA ARCHIVE (COVID-ARC) powered by the Laboratory of Neuro Imaging (LONI). For up-to-date information on the study, visit https://covid-arc.loni.usc.edu/. COVID-ARC is funded by the National Science Foundation (NSF) under award number 2027456.

1.2 Research Objectives

Federated learning offers a novel approach to health diagnostics which preserves the privacy of the subjects of the underlying data. The COVID-19 pandemic has increased the availability of data to be used in federated learning architectures. This research explores the feasibility of using federated learning and CNNs for the detection of COVID-19 in chest CT scans. The research contributes to the literature in three ways: firstly, it makes use of a robust and extensive base of sites and training data which replicates real-world circumstances; secondly, the experimental design which compares a range of models and training regimes for different sites reveals important findings for the conditions necessary for achieving high levels of accuracy; thirdly, it explores the capacity for these models to generalise to data from sites not included in training, and evaluates whether or not these models are sufficiently accurate to be used as an alternative or supplementary diagnostic tool by healthcare facilities that did not participate in training.

The third component has important implications for public health and is the focus of this research - a model that can generalise to sites that were unable or unwilling to contribute training data could leverage this model for their own diagnostic activities. This is of particular importance for lower-resourced sites that may not have the infrastructure, skills or other pre-requisites needed to participate in training. It may likewise be of importance to facilities that are over-utilised, or are unwilling to contribute data. In all cases, these models may alleviate pressure on radiologists, or offer superior results to and avoid the costs of each facility developing their own model.

Research into federated learning and diagnostics rarely explores whether and how these models are practically useful for facilities that did not participate in training, and instead focus on understanding the ability of the model to generalise to the test data of the sites that contributed training data. This research compares the test accuracy of federated models for sites that did not participate in training the federated model to the test accuracies of independent models developed by these same sites. Accuracy would be the key metric that a facility decision-maker would evaluate when considering whether or not to use the federated model as an alternative or supplementary diagnostic tool, or to develop and use their own model - the decision-maker would choose to use the model with the higher accuracy as they are unable to control the balance of the underlying training data and the balance of the data of scanned patients that present for diagnostics. Evaluating these models from the perspective of the decision-maker or 'user' is a further unique contribution of this paper.

The findings can help to direct further research into how federated learning for diagnostics can be used by sites that are not participants in training. Achieving this would support lower-resourced or over-utilised sites and widen access to modern diagnostic capabilities.

1.3 Background

The novel coronavirus 2019 (2019-nCoV, or COVID-19) is a communicable respiratory disease that was first detected in Wuhan City in China, and first reported to the World Health Organisation (WHO) on the 31st of December, 2019. The disease is highly infectious and is spread primarily through airborne transmission and to a lesser extent, transmission through physical contact. COVID-19 typically presents through respiratory manifestations similar to the flu such as fever, cough and fatigue, and may present with systemic manifestations such as multiple organ dysfunction, stroke, or confusion [Bell 2020].

The disease spread rapidly beyond Chinese borders until being declared a global pandemic on the 11th of March, 2020. Countries around the globe were subject to 'waves' of infection of varying intensity leading to a substantial number of lives lost. The first infectious waves caused significant and widespread social and economic challenges. Public healthcare infrastructure and resources were in many cases unable to manage the increased demand on the healthcare system. In some countries, healthcare systems were overwhelmed. Amongst others, this was driven by shortages of ICU beds, ventilators for treatment, and polymerase chain reaction (PCR) testing kits for diagnostics. The impact of COVID-19 could therefore have been greatest in countries with pre-existing shortages of medical infrastructure, such as the Global South.

The crisis and associated pressures stimulated a wide range of research into innovative solutions to help mitigate and resolve the crisis. Artificial Intelligence (AI) applications were explored and tested for a variety of use-cases. However many of these use cases rely on access to personal health information which is often subject to special sharing and use conditions contained in data protection legislation.

1.3.1 COVID-19 and AI

The pandemic stimulated a wide range of research exploring how AI can support efforts to control infections and mitigate their impact on the health sector and society. Arora *et al.* [2020] provide an overview of the application of AI in COVID-19:

- **Prediction and Tracking:** historical data on infections, and socio-economic and demographic data can be used to predict infections, the timing and scale of an infectious wave, and the expected period until a wave subsides.
- **Monitoring of COVID-19 Cases:** vital statistics and clinical parameters for patients admitted to health-care facilities can be used to identify an optimal course of treatment and prioritise resource allocation amongst patients for ventilators and respiratory support systems.
- **Early Diagnosis:** chest X-Rays and CT scans can be used to diagnose whether an individual has COVID-19 in the absence of PCR testing kits or other diagnostic facilities.

- **Protein Structure Prediction:** the structure of proteins needed for virus entry and replication can be identified through AI. This provides insights that support drug development.
- **Developing Therapeutics:** lead discovery, virtual screening and validation processes for the development of new therapeutics can be accelerated by AI through the assessment of the molecular descriptors and properties of approved and validated drugs that could be repurposed for COVID-19.
- **Developing Vaccines:** the development of vaccines could be accelerated by AI systems that predict potential vaccine candidates.
- **Curbing the Spread of Misinformation:** extensive information on the pandemic is accessible, however, information on the pandemic varies in its authenticity, accuracy and integrity. AI can be used to classify whether digital content is 'fake' to curtail its propagation. It can also be used to understand trends in public sentiment, and identify important COVID-19 related topics people are discussing on social media.
- **Genomics:** AI can be used to quickly classify COVID-19 genomes through the analysis of genomic structures.

1.3.2 Data Protection and Healthcare

Countries around the world are rapidly implementing data protection legislation. UNCTAD [2020] estimates that 80% of countries have implemented legislation or have draft legislation. Countries in the developing world are those least likely to have current or draft legislation however this is expected to change as digital rights are increasingly recognized as fundamental human rights, and the absence of protection of personal data inhibits cross-border trade in digital services which are an increasingly important channel for economic development.

The General Data Protection Regulation (GDPR) is arguably the global standard of data protection and a model that many countries adapt to their local conditions [EU 2022]. The GDPR embraces six principles of data protection that empower individuals with control of their personal data. These principles include: lawfulness, fairness and transparency; purpose limitation; data minimisation; accuracy; storage limitation, and; integrity and confidentiality. These principles collectively determine the conditions under which personal data can be collected, stored and used and the responsibilities of data processors and collectors when collecting, storing and using personal data.

These principles or adaptations thereof are therefore common across the globe. GDPR and many local adaptations of the legislation consider health data especially sensitive and classify it as 'special category' or 'special person' data. This class of data often requires explicit consent for its use and limits its use in processing to health-related purposes only.

These limits and expectations create scarcity in the availability of rich and complete datasets for healthcare research. These limits and expectations can be severe – for

example, it was reported that Google withdrew the release of 100,000 anonymized chest x-rays of UK citizens because details such as the date of the x-ray and jewelry in the x-rays may enable reidentification [MacMillan and Bensinger 2019]. Researchers therefore require technical solutions that can support analysis of personal data while allowing them to meet domestic and international data protections obligations.

CT scans presented an important alternative to diagnosing COVID-19 while managing a shortage of traditional diagnostic services. Despite this, openly accessible CT scan data for research are fragmented and in short supply. Centralised data sources necessary for research and the development of healthcare solutions are often consolidated by research centres with restricted access. As an alternative, some researchers have sought to collate data from a multitude of openly-accessible sources to support diagnostic research and overcome scarce training data.

While many of these collation undertakings are commendable, they suffer from a range of shortfalls such as: dependency on 'composite data-sets' where images have been extracted from a wide range of research papers; scarce features in associated meta-data preventing segmentation of the images by location, age, or other important features; and, heterogeneous data quality. Although these datasets allow for research and experimentation in machine learning driven diagnostics, they do not adequately reflect the complexities and limits associated with developing diagnostic solutions using special person data. These datasets are also unlikely to meet the standards needed to support high-quality clinical research that can be used in production.

1.3.3 Federated Learning, Transfer Learning and Data Scarcity

Traditional machine learning models require all data to be centralised for training and evaluation. Federated learning is a privacy preserving machine learning technique that does not require the centralisation of data. This enables the training of a machine learning model using data from a variety of sources without requiring the data be shared between these sources. Federated learning relies on a 'shared model' that is updated from the weights or gradients from epochs learned by 'local' models that are each trained using an independent dataset as input. The training data used in the local models are therefore not shared with the custodian of the shared model nor is it shared with the custodians of each of the local models. This creates a shared model that is trained using all available data, and that can then be used for prediction by a variety of participants. A more detailed explanation of federated learning is discussed in the literature review.

The federated learning approach was developed by Google in 2017 McMahan *et al.* [2017]. Federated learning was first used to support next-word prediction on Gboard on Android. This is the Google Keyboard found on many mobile devices. A federated learning approach enabled training for next-word prediction on local devices by using data from the device owner's interaction with their keyboard without having to export sensitive user data to shared servers.

The ability to train machine learning models without having to share data makes federated learning particularly attractive for the health sector. Rieke *et al.* [2020] highlight

how barriers to healthcare research incurred by data protection obligations can be overcome by federated learning. The authors highlight that the need for centralised data for healthcare research not only poses regulatory, ethical and legal challenges, but also technical challenges such as the need for anonymisation, access control and safe transfer of data which in some cases can also prevent research.

Federated learning enables healthcare institutions to undertake research and develop healthcare solutions without being dependent on access to data from research centres or from lower quality datasets composed from a variety of public sources. Federated learning is therefore considered a potentially 'disruptive' technology which empowers data custodians to define the conditions under which data is used in training and validation. The technology is noted as having application in identifying clinically similar patients, and predicting hospitalisations due to cardiac events, mortality and ICU stay time [Rieke *et al.* 2020]. Rieke *et al.* [2020] furthermore highlight that the COVID-19 pandemic has stimulated research in the intersection between federated learning and healthcare. Research in this field could have applications outside of COVID-19.

Transfer Learning and Data Scarcity

Federated learning can increase the availability of data needed to train machine learning models. The challenges of scarce training data can also be mitigated through transfer learning. Transfer learning is the process of 'fine-tuning' a machine learning model that has been pre-trained using another dataset. Transfer learning is often used in computer vision and machine learning for image recognition tasks. Transfer learning can be used in conjunction with federated learning.

'ImageNet' illustrates the process and utility of transfer learning [Deng *et al.* 2009]. ImageNet is a database of 14 million annotated images which can be used to train image classification models. Transfer learning is often conducted using this database, where image classification models are firstly trained using this database, and then retrained using a smaller, task-specific dataset. Pre-training establishes a model that is adapted to image classification generally and is inherently able to extract common features such as edges and shapes, while retraining optimises the model to classify images for the given task [Heidari *et al.* 2020].

In the context of deep learning, transfer learning establishes a network with strong feature extraction capabilities learned from the larger dataset which can be oriented to a specific task without the need for as substantial a training set as ImageNet, for example. Transfer learning furthermore mitigates some of the risks of overfitting that may arise when training the network with a small training dataset [Heidari *et al.* 2020]. Transfer learning is well suited to training custom deep learning networks [Masud 2022].

Chapter 2

Literature Review

The following provides a review of existing literature related to the research. There are four subsections. The first subsection provides an overview of CNNs used in image classification tasks, VGG-16, and existing research in their application in COVID-19 diagnostics. The second subsection provides an overview of federated learning. The third subsection provides an overview of the use of CNNs in federated learning for COVID-19 diagnostics. A summary of gaps in the research is presented in the final subsection.

2.1 Convolutional Neural Networks and VGG16

CNNs are deep-learning neural networks that allow for the processing of arrays such as images. The field is continuously evolving as researchers explore new architectures and use cases. CNNs often have common components across different architectures. VGG16 is a frequently used CNN and part of the VGG family of CNN.

VGG16 uses a sequence of convolution and pooling to reduce the dimensionality of the data while maintaining important features. Figure 2.1 details the VGG16 architecture. Important components of this architecture are discussed below. VGG16 models pre-trained on ImageNet are publicly available.

Input layer: the input layer is equivalent to the dimensions of the input image. The VGG family of CNN often takes as input an image of 224 x 224 x 3. The input is not flattened as would be needed for a standard feed-forward neural network. This allows the network to learn features reflecting the spatial and contextual dependencies between pixels.

Kernel Filters & Convolution: The input is processed through 5 sets of convolution layers. Each set has 2 or 3 convolution layers with 3×3 with 1 stride kernel filter to convolve the image. This structure preserves the spatial resolution of the data. The outputs of each convolution layer are passed through a 'Rectified Linear Unit' or 'ReLU' activation function. A ReLU activation replaces any negative value to 0 which improves the performance of the network and prevents the risk of vanishing gradients. The

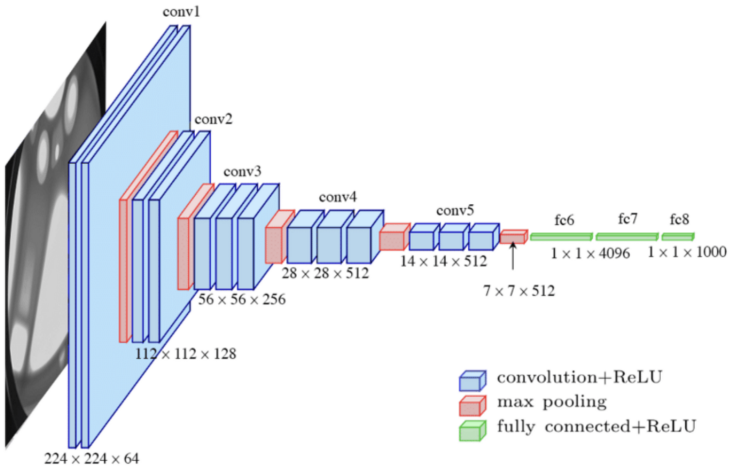


Figure 2.1: The VGG16 Architecture [Ferguson *et al.* 2018]

following describes the ReLU layer where (x) is a neuron input. The convolution and activation therefore extract important features in the data.

$$R(x) = \max(0, x).$$

Pooling: The output of each convolution layer is traditionally passed through a pooling layer of 2x2 with stride 2. This reduces the dimensionality of the data by halving the size of the activation. Max and average pooling are common transformations conducted in the pooling activity - max pooling retains the largest activation from the kernel whereas average pooling creates an output with the average of the values from the kernel. Max pooling therefore retains the most prominent features from the input, whereas average pooling retains the average of features in the image.

The VGG family has 5 sets of convolution and pooling, however models in the family vary in the number of convolution layers each convolution set contains. Differences in architecture are denoted by the number in the title of the network (e.g. VGG16 has a total of 16 layers with weights, of which 13 are convolutional layers and three are dense layers). This is highlighted in Figure 2.2. Transfer learning with VGG16 is typically conducted by retraining the dense layers - earlier convolutional layer sets are adapted to extracting foundational features such as shapes, edges and colours while later features and the dense layers detect more complex and task specific features such as eyes or lesions in chest CT scans.

Dense Layers and Output: following the 5 sets of convolution and pooling, the output is flattened as an input into two fully-connected dense layers. In a standard VGG16 each of the first two dense layers have 4,096 neurons. The output of the last dense layer is passed through a final activation layer for classification which corresponds to the number of categories for classification which is 1,000 in the standard VGG16. A softmax activation function is often used, highlighted below. This activation function indicates the normalized probability that an input is a given class. Classification is therefore made possible by the process of 'encoding' through convolution and pooling where the image's dimensionality is reduced.

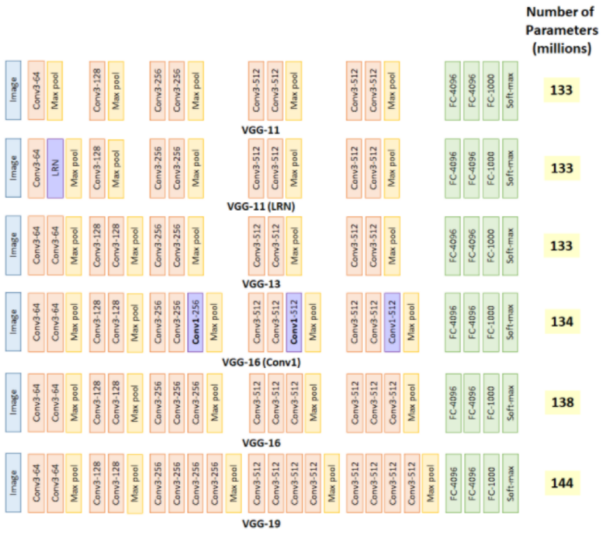


Figure 2.2: The VGG family of CNN [Tsang 2014]

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

$$\begin{aligned}
\sigma &= \text{softmax} \\
\vec{z} &= \text{input vector} \\
e^{z_i} &= \text{standard exponential function for input vector} \\
K &= \text{number of classes in the multi-class classifier} \\
e^{z_j} &= \text{standard exponential function for output vector} \\
e^{z_j} &= \text{standard exponential function for output vector}
\end{aligned}$$

Dropout: VGG16 can be augmented through the introduction of drop-out layers between other connected layers. A drop-out layer allocates each neuron a probability of remaining active or being set to 0 for each step. Drop-out can limit overfitting and improve generalisation.

Optimizers and Learning Rates: VGG16 is trained through backpropagation which facilitates the update of weights in the network. This is achieved by comparing the network's predicted values with the actual values to calculate the network's error. An optimizer is then used to minimize the cost function of the network based on this error, which is then used to update network weights. A learning rate is used to control the size of the updates to the network: learning rates that are too small may result in slow training while learning rates that are too high may result in fluctuations in network accuracy and an inability for the network to converge to the global minimum. Optimizers

vary in their structure Kandel *et al.* [2020]: vanilla gradient descent optimizers such as stochastic-gradient descent (SGD), batch gradient descent, and mini-batch gradient descent; momentum-based gradient descent such as Momentum Gradient Descent and Nesterov Momentum Gradient Descent, and; adaptive gradient descent such as Adam, RMSProp, and AdaGrad. Kandel and Castelli [2020] evaluate the impact of varying the learning rate and optimiser on the accuracy of VGG16. The authors find that the lowest learning rate in conjunction with Adam yields the greatest accuracy and fastest convergence.

VGG Network Customisation and Transfer Learning: the VGG family of CNN are well suited to customisation to best achieve a given task. For example, this might include adjusting the size of the input layer to facilitate processing of images of different sizes, or adjusting the size and number of dense layers. Decreasing the latter can improve training time. In some tasks this can be done without significant deterioration in results. VGG16 networks that are pretrained on datasets such as ImageNet are publicly accessible. Transfer learning is therefore readily possible and facilitates fine-tuning a model with custom dense layers, or other customisations.

Research into the detection of COVID-19 through feature extraction of x-rays and CT-scans has accelerated. Salman *et al.* [2020] appear to be one of the first published efforts for using CNNs in the detection of COVID-19 with chest x-rays. The authors made use of a pre-trained InceptionV3 and achieved 100% accuracy - comparable to expert radiologists. The transfer learning approach catered for the small dataset which contains 130 COVID-19 and 130 Normal X-ray images collated from a range of sources. Das *et al.* [2020] also appears to be one of the first published papers in the field. The authors made use of Xception with a Softmax activation function. The Xception model was pre-trained on a larger dataset to allow fine-tuning with the smaller COVID-19 x-ray dataset. The authors compared their proposed approach to a range of CNN and other machine learning classification techniques where they found their approach as the most accurate at 97%.

Tsiknakis *et al.* [2020] highlight the noted importance of transfer learning to detect COVID-19. The authors trained InceptionV3 on the ImageNet dataset. The pre-trained model was then fine-tuned with the COVID-19 training data where the 'backbone' layers were frozen. A total of 522 images across four classes were used with confirmed COVID-19 images accounting for 25% of the dataset. The authors achieved better performance than comparable studies in the binary classification of COVID-19. The model's performance was inferior to comparable studies in the ternary and quaternary classifications.

The comparable performance of the range of CNN architectures available for COVID-19 diagnostics was investigated by [Haque *et al.* 2021]. The authors compared five pre-trained CNNs and included weather variables as inputs. Models evaluated includes VGG16, VGG19, Xception, InceptionV3 and Resnet50. The models made use of a Soft-Max activation function to classify images into 1 of 3 groups. The authors used a categorical cross-entropy loss function and ADAM optimizer. The VGG16 and VGG19 models demonstrated the highest accuracy after training. Both achieved over 92% accuracy for the validation set.

Kassania *et al.* [2021] also compared model performance to identify the optimal pairing of a CNN feature extractor and a machine learning classifier. The authors run a battery of experiments with 8 CNNs and 6 machine learning classifiers. The CNNs evaluated include: MobileNet; DenseNet; Xception; ResNet; InceptionV3; Inception-ResNetV2; VGGNet; and NASNet. The machine learning classifiers include: Decision Tree; Random Forest; XGBoost; AdaBoost; Bagging Classifier; and LightGBM. A total of 137 images were used. No data augmentation was undertaken. The authors found the combination of a DenseNet121 feature extractor with Bagging tree classifier had the highest accuracy at 99%.

Fibriani *et al.* [2020] demonstrated that leveraging a range of models collectively in a majority voting approach using a set of CNNs yields superior results. The authors create a system of nine different CNNs which are combined using three different majority voting techniques. The CNNs the authors tested includes: AlexNet, ResNet 50, Wide ResNet 50 – 2, ResNet 101, ResNet 152, VGG16, VGG16 Batch, Normalization (BN), VGG19, and VGG19 Batch Normalization (BN). The 'Soft Majority Vote' model achieved the highest accuracy at 99.3%. The soft majority vote model made use of a weighted majority voting technique.

The use of novel architectures and optimisation of architectures for the diagnostic task has a less pronounced literature. For example, Afshar *et al.* [2020] aim to demonstrate that a capsule network pre-trained for classifying thorax diseases can overcome the limitations CNNs incur when using small training datasets while delivering superior accuracy. Mukherjee *et al.* [2021] proposed a nine layer CNN to illustrate that CNNs trained with both chest x-rays and CT scans will more accurately classify COVID-19 as the training allows the model to learn more anomalous patterns caused by COVID-19. The proposed model achieves a 96.3% accuracy compared to 71.9% for InceptionV3, 82.9% for MobileNet, and 90.8% for ResNet.

Customised networks are able to achieve comparable results to vanilla architectures of well established networks such as VGG16 and ResNet. Customisation is often undertaken by varying the number of layers, the size of these layers, or introducing a dropout function. Yang *et al.* [2021a] illustrate that customised VGG16, DenseNet121,ResNet50, and ResNet152 can yield superior results for the classification of COVID-19 using chest X-Rays and CT scans. Customisation is undertaken by altering the shape and size of dense layers. For example the VGG16 architecture achieves 98% accuracy with four dense layers being used before the final output layer with the following shapes: 1024, 512, 256 and 128. Kumar *et al.* [2022] developed a Modified-VGG16, Modified-VGG19, Modified-ResNet50, and Modified-InceptionV3 to diagnose COVID-19 using chest X-Rays. These networks were pretrained in ImageNet and customised to have a single dense layer of size 64 before classification. All models achive superior accuracy to their non-customised counterparts.

There is variation in the approach to image pre-processing before training. Approaches to pre-processing CT imagery for deep learning are discussed in Chapter 3. Heidari *et al.* [2020] illustrate that removing diaphragms and undertaking image noise filtering and histogram equalization improves the test accuracy of a VGG16 model from 87.6%

to 91.0%. This prevents the model from being exposed to irrelevant areas of the chest, and supports detection of important features in relevant areas of the chest.

2.2 Federated Learning

Federated learning is a privacy preserving approach for developing machine learning models trained using a variety of data sources. The approach relies on the transfer of gradients or weights between a shared model and local models. Figure 2.3 provides a high-level overview of the structure of a federated learning project[1].

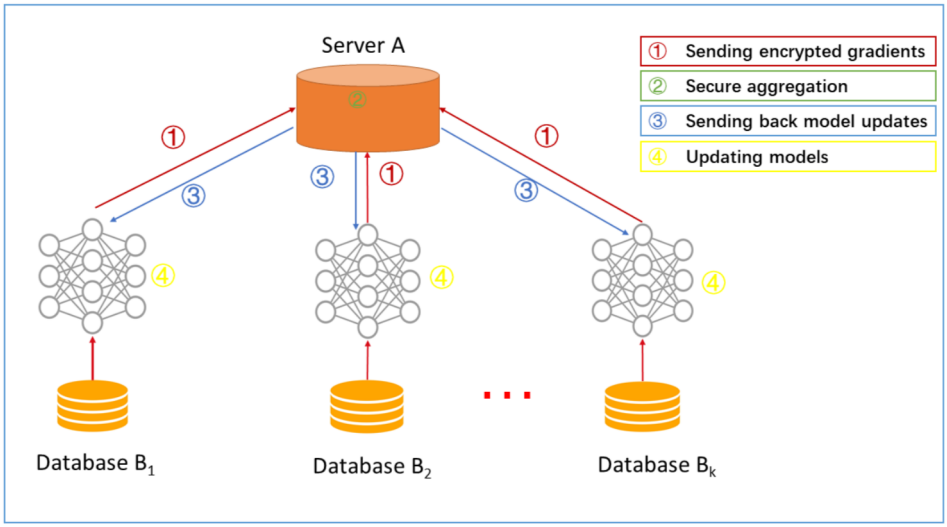


Figure 2.3: A Stylised Federated Learning Network [Yang *et al.* [2019a]]

There are two categories of FL [Yang *et al.* 2019b]. The first category of FL is vertical FL where the datasets of the local models share common unique identifiers but do not have equivalent features. For example, this might include data held between two bank units where the units have data for a common set of individuals however the first unit has individual spending data and the second unit has individual borrowing data. The second category of FL is horizontal FL where the datasets of the local models have equivalent features. For example, this includes a set of hospitals that each service different patients but each have patient x-rays and data on the age and sex of these patients. Horizontal FL is the focus of this research.

Federated learning architectures differ in how the local and shared models interact and update. Three of these are explained below [Guo 2020]:

- **Federated Stochastic Gradient Descent (FedSGD)** refers to the process of training a shared model where the shared model's gradient descent step is calculated from the average gradients from steps taken by the local models.

[1]The diagram highlights a Federated Stochastic Gradient Descent approach to FL

- **Federated Averaging (FedAvg)** refers to the process of training a shared model where the weights of the shared model are updated to the average weights of the local models following an epoch of local model steps. FedAvg is approached by ensuring the local models are initialized with equivalent weights.

- **FedProx** refers to the process of training a shared model where local models can submit partial work which allows for differences in the extent to which each local model is trained before submitting weights to the shared model. The extent to which the work is partial determines the weight of the local model's weights in update.

Guo [2020] highlight that the speed to convergence and accuracy between FedProx and FedAvg is somewhat unpredictable when using non-identically and independently distributed (NIID) training data, where some tests indicate FedAvg offers superior speed to convergence. Brendan McMahan *et al.* [2016] proposed the FedAvg algorithm and evaluated it relative to FedSGD. The authors demonstrate that the algorithm achieves comparable accuracy and superior speed to convergence relative to FedSGD. This research therefore uses a FedAvg approach, which also appears to be common in the literature in the use of federated learning for the detection of COVID-19 in chest CT scans and X-Rays.

The training process of FedAvg usually contains the following steps [Yang *et al.* 2019b]:

1. A shared model is elected;

2. Shared and local models are initialized with equivalent weights;

3. An epoch of a predefined number of steps allows each local model to update its weights;

4. The weights of the local models are sent to the server for the shared model. The local model weights are averaged and used to update the weights of the shared model, and;

5. The updated weights from the federated model are transmitted back to the local models where each local model updates its weights.

Steps 3 to 5 repeat until the loss function of the shared model converges.

Abdul Salam *et al.* [2021] compare the accuracy of federated average and traditional machine learning approaches to diagnosing COVID-19 with flattened chest x-rays as inputs. Both used a sequential deep learning model. The authors experimented with SGD and ADAM optimizers, and binary cross-entropy and sparse categorical cross-entropy loss functions. The authors also varied the learning rates. The FL model with SGD algorithm had a higher prediction accuracy, higher training time and a lower prediction loss than the traditional model. The softmax activation function and SGD optimizer provided better prediction accuracy and loss.

2.3 Federated Learning and CNN for Diagnostics

The preceding section highlights that the first step in a FL architecture is the election of an appropriate model. A CNN can be integrated in a FL architecture to support training for processing of images. Figure 2.4 provides an overview of the use of FL in health using a CNN. The diagram draws from a paper that made use of an FL approach for the segmentation of chest CT scans to detect COVID-19. The diagram highlights the use of three local models with data from independent hospitals that are used to train a shared model.

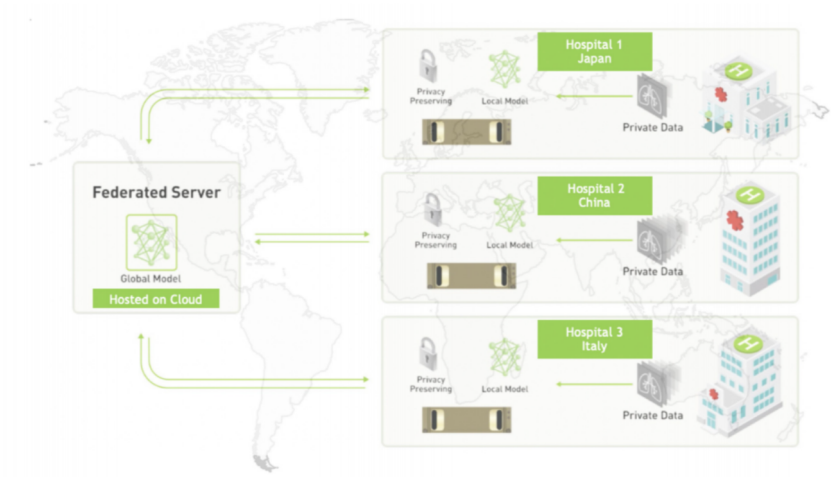


Figure 2.4: A Federated Learning Approach to COVID-19 Diagnostics [Yang *et al.* 2021b]

There is a small but growing base of research into FL and CNN for the detection of COVID-19. Liu *et al.* [2020] appear to be one of the first published papers on the use of FL in the detection of COVID-19 using x-rays. The authors compared the performance of federated averaging and traditional machine learning. The authors compare four networks: CovidNet, ResNeXt, MobileNet-v2, and ResNet18. All networks in the FL model converged slower and to a higher loss than without FL. The authors also found that ResNet18 has the fastest convergence speed and highest accuracy. The authors do not provide clarity on how the images are segmented in the FL model. The authors note that they define a set of image owners, allocate images to the owners and assume that the owners are different hospitals. This appears similar to randomly splitting images into different 'owners'.

Yang *et al.* [2021b] also appears to be one of the first published papers on the use of FL in the detection of COVID-19 using CT scans. The authors make use of a weighted FedAVG and a U-shaped fully convolutional network. The authors also implemented a semi-supervised learning approach to using training data from sites without annotations. The authors highlight that models trained with data from a single site do not generalise as well as those trained with multiple sites. The authors also highlight that protection of personal information has inhibited the sharing of COVID-19 data motivating the need for FL. CT scans from different sites vary in the observable patterns that

indicate the existence of COVID-19. This includes size, shape and texture. Differences in the distributions of data between these sites results in imperfect generalization of models built using a single site's data. The research provides an important foundation for analysis of FL and CNN for the detection of COVID-19. However, the study has a narrow range of sites. Furthermore, one of the sites does not have annotated data.

He *et al.* [2020] present a novel AutoFL model – FedNAS (Federate Neural Architecture Search) – which improves the accuracy of FL by searching for architectures that deliver better results. The authors vary the hyperparameters and the CNN network architecture. The study illustrates that FedNAS yields higher accuracy in the classification of the CIFAR10 dataset than FedAvg (81.2% vs 77.8%). The authors highlight that the findings will be useful in diagnostics for COVID-19 using CT images however do not evaluate their model with this use case.

A further novel approach is presented by Zhang *et al.* [2021]. The authors make use of GhostNet, ResNet50, and ResNet101 for the classification of COVID-19. The authors present 'a dynamic fusion method' for the update of the shared model. The method determines which local models participate in a given shared model update based on the performance of the local models, the local models' training schedules and the local model' training times. The dynamic fusion method achieved a faster rate of convergence and a higher accuracy. The research paper provides a useful overview of the opportunity for FL in the detection of COVID-19. However, it appears that the input data does not contain site identifiers and that the authors constructed clients and then randomly allocated images to each client. It is therefore unlikely that there is any meaningful variation in image sets between different clients.

2.4 Summary of Existing Research

Review of the literature highlights that investigation into the use of CNN and FL for the detection of COVID-19 is progressing and is revealing important complexities in developing usable and accurate networks for COVID-19 diagnostics[2]. The literature review highlights some key learnings for the approach to the research: VGG16 is an attractive architecture; a softmax activation function performs well for classification; transfer learning using ImageNet is advisable; customisation of networks can yield superior results; image reshaping and normalisation are minimum pre-requisites for pre-processing. These learnings are adopted in this research.

The review illustrates how this research contributes to the literature. There is a rich base of research that highlights how CNNs can be used to diagnose COVID-19 using chest x-rays and CT scans and a narrower base of research on the use of FL to support COVID-19 response and diagnostics. While this base is expanding, researchers are not

[2]The table in the appendix provides a summary of literature relevant to the research. The table provides an overview of the models to highlight appropriate designs and approaches. The table also highlights the objective of the study, findings from the study and key learnings for the proposed research

yet exploring the accuracy of FL in diagnostics using data from sites not included in the training dataset - models are evaluated from test data from sites included in training. This highlights a gap in exploring the practical utility of federated learning in healthcare from the perspective of a facility deciding whether or not to use the federated model. Existing research often suffers from further limitations: firstly, most of the research has unclear segmentation of images into sites – a site-level indicator in the data that would facilitate a real-world replication of the conditions for collectively training a diagnostic system is often unavailable; secondly, papers that have a clear and robust approach to segmenting images between sites typically use a small number of samples from a limited number of sites.

Chapter 3

Data and Methodology

This chapter details the data and methodology used in this research. There are 2 sections. The first section explores the source dataset through descriptive statistics and samples, and discusses common approaches to pre-processing undertaken by research with a similar task. The second section explains the methodology undertaken in this research. This includes a discussion on data pre-processing and the customised VGG-16 model. It also includes the experimental training and testing regime, and associated hyper-parameters.

3.1 Data

The research uses a novel data source which has collated CT scan and X-ray data for detected COVID-19 patients and healthy patients. The 'COVID-ARC' data is segmented by site. The COVID-ARC data solves for the shortfalls of other collated data sources that lack site-level segmentations. Table 3.1 provides an overview of relevant data that has been made available for this research. The COVID-ARC repository provides access to CT scan data from 15 sites. These sites are distributed across the globe which creates the potential for geographically relevant variation in data between sites. This variation can be exploited to create more robust and accurate diagnostic models than models built using a single site. The wide number of sites could also enable generalisation to sites not included in training. The data is presented in six different formats.

A total of 278,713 CT scans indicate positive COVID-19 cases. This data is unevenly spread across sites. A total of 312,331 scans indicate negative COVID-19 cases. These negative COVID-19 cases include healthy patients, and patients with a detectable respiratory disease that is not COVID-19. Only 5 of the 15 sets offer both classes of data. There is imbalance between the positive and negative COVID-19 data for these five sites.

Positive COVID-19 cases are detectable in CT scans through a variety of observable traits. This includes lesions, ground-glass opacity (GGO), crazy-paving pattern, and vascular dilation. Chun-Shuang Guan and Chen [2020] highlight that round GGOs tended to evolve into patchy GGOs through the course of infection with the virus. These

Table 3.1: Source Data Overview

Site Indicator	Location	Modality	File Format	No. of Images		Number of Patients	
				COVID-19	Non COVID-19	COVID-19 Patients	Non COVID-19 Patients
1	Sao Paulo, Brazil	CT	.png	2,168	2,005	80	130
2	China	CT	.png	349	463	216	55
3	Italy	CT	.nii	110	-	60	-
4	Wenzhou, China and Netanya, Israel	CT	.nii	20	-	-	-
5	Valencian Region Medical Bank	CT	.tgz	900	-	-	-
6	Moscow, Russia	CT	.nii	1,110	-	1,110	-
7a	SAR, Iran (Total)	CT	.tif	15,589	48,260	95	282
7b	SAR, Iran (For Machine Learning)	CT	.tif	2,282	9,776	95	282
8	Various	CT	.png	84	-	-	-
9	University of Electronic Science and Technology, China	CT	.nii	120	-	120	-
10	China Consortium of Chest CT Image	CT	.png	159,702	251,827	408	408
17	Radiology Society of North America	CT	.nii	632	-	632	-
18	Arkansas, USA	CT	.dcm	31,935	-	105	-
19	Radiology Society of North America	CT	.dcm	31,856	-	110	-
25	Catanzaro, Italy	CT	.nrrd	31,856	-	50	-
TOTAL	-	-	-	**278,713**	**312,331**	**3,081**	**1,157**

presentations of COVID-19 can therefore be detected through CNNs. Figure 3.1 is an annotated sample of chest CT scans highlight some of the most common observable manifestations of COVID-19 in chest CT scans [Chatzitofis *et al.* 2021].

The research was undertaken using the 4 sites with COVID-19 and non-COVID-19 images. A summary of this data is highlighted in Table 3.2. Site 10 was manually decomposed into two sub-sites. These sub-sites each have observably similar images and have equivalent file-types which enabled their segmentation. The research was undertaken

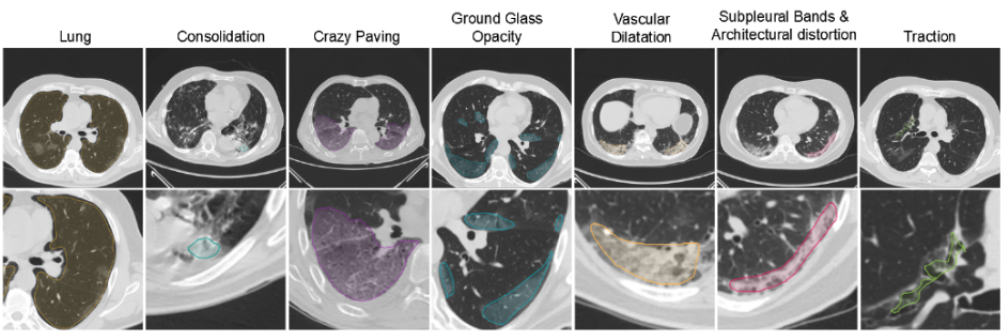


Figure 3.1: Lesion annotations of chest CT scans for COVID-19 positive patients

using the 5 sites with COVID-19 and non-COVID-19 images. A summary of this data is highlighted in Table 3.2.

Figure 3.2 is a sample of the images from the elected sites. The top row contains positive COVID-19 cases and the bottom row contains negative COVID-19 cases.

The sample highlights significant variation in the underlying structure and content of the images between sites. This is attributed to differences in underlying CT scan infrastructure and approaches to image processing before submission of the imagery to the COVID-ARC repository. For example, Site 2, has masked non-lung content.

Table 3.2: Input Data Used in Training

Site	Model N (Positive; Negative)	Shape	Range
Site 1	2,926 (2,168; 758)	Multiple ($\sim 365 \times 260$)	[0, 1]
Site 2	747 (397; 350)	Multiple (wide ranging, limited consistency)	[30, 250]
Site 7b	9,961 (2,283; 7,633)	512×512	[0, 1]
Site 10_1 (png)	20,000 (10,000; 10,000)	512×512	[0.17, 1]
Site 10_2 (jpg)	14,310 (4,310; 10,000)	512×512	[0, 255]
TOTAL	47,944 (19,158; 28,741)		

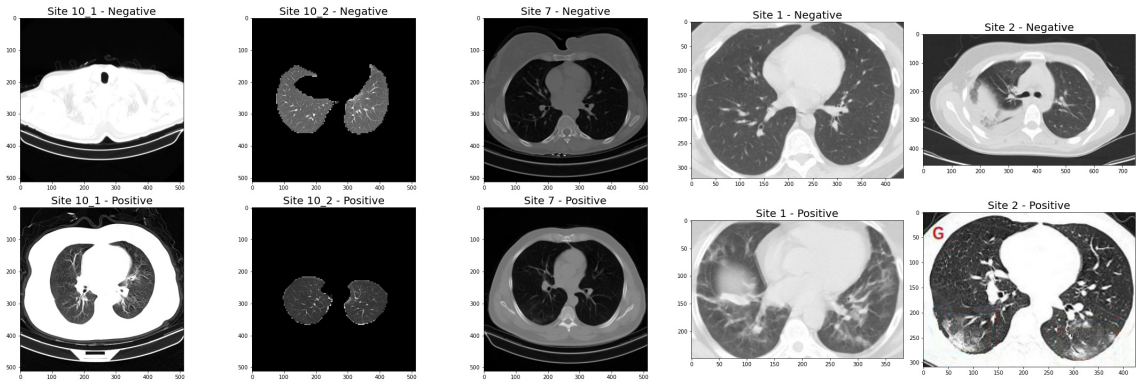


Figure 3.2: Sample of Site Images Used in Training

The sites therefore vary in terms of their compliance with standard practices in the pre-processing of CT imagery for machine learning. Masoudi *et al.* [2021] detail the range of preprocessing steps often used to prepare CT imagery for deep learning. The steps used and how they are implemented are determined by the research objectives and nature of the underlying data. This includes:

1. **Denoising** where the image is transformed in spatial and frequency domains. This is done to cater for disturbances in the imagery which may be caused by, "beam hardening, patient movements, scanner malfunction, low resolution, intrinsic low dose radiation, and metal implants" Masoudi *et al.* [2021].

2. **Interpolation** where the image is transformed to establish consistent image resolutions. This is done to ensure alignment between the input layer and the image dimensions. Masoudi *et al.* [2021] highlight that cubic-spline and B-spline are preferred convolving functions. Tan *et al.* [2021] demonstrates that downscaling the resolution of input images can deteriorate results. The authors highlight that upscaling chest CT scans for the classification of COVID-19 using a VGG-16 improve results. They achieve this through the use of an SRGAN neural network to construct higher resolution images. Models trained using these images achieve superior precision, recall and accuracy than their lower resolution counterparts. This suggests that maintaining image resolutions or upscaling is preferable to downscaling.

3. **Registration** where the image is transformed to have similar placement of areas of interest. For example, this would be done to ensure the chest is centred in

the CT scan. This is done to support the network's ability to learn and extract important features from the data by creating consistency in inputs.

4. **Windowing** where the image contrast is increased in regions of interest. This is done to increase variation in areas of interest. This can improve the network's sensitivity to the variation in input images' areas of interest.

5. **Normalisation** where the image's range of values is standardised. For example, remapping image values to the range of [0,1] or [0, 255]. This is done to provide the network with consistent inputs and can be conducted at the patient or institution level.

Figure 3.3 is an additional sub-sample of Site 2 - the composite site. This demonstrates the significant variation in image structure of the composite site. This site highlights the real-world complexity of using data from a range of institutions when developing collaborative deep learning solutions to CT diagnostics. These images vary in their resolution and range.

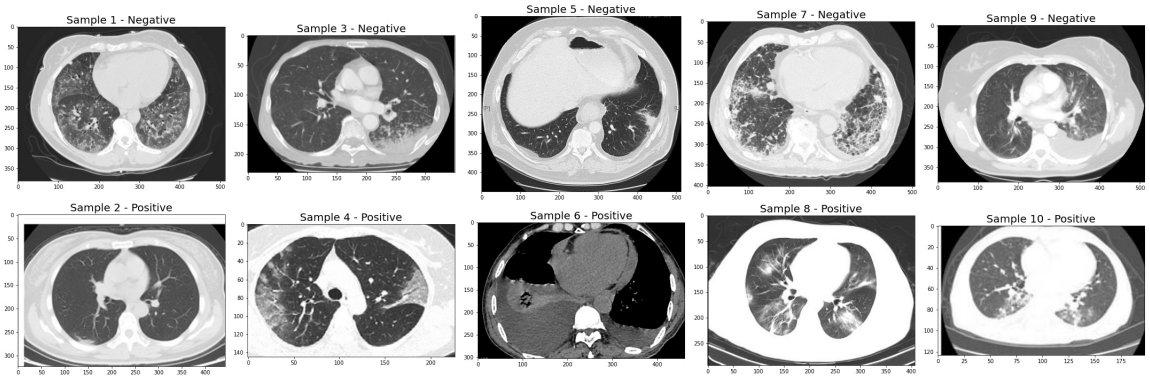


Figure 3.3: Sample of Site 2 Images

3.2 Methodology

The methodology has been designed to explore whether a deep learning network can generalise to data from sites not used in training. This informs the approach to data preparation, the model architecture and its hyperparameters, as well as the training regime. The methodology was also developed through review of existing literature. The methodology uses consistent data preprocessing across sites, and equivalent models and training regimes to ensure comparability of inputs and results.

3.2.1 Data Preparation

All input data was pre-processed using a standardized pipeline for all sites. This pipeline ensured all images were equivalent in range and dimensionality. There are two steps in the pipeline:

1. **Normalisation** where all images were transformed to lie within the range [0,1]. The range is calculated for each image and not from the entire dataset.
2. **Interpolation** where all images were scaled to have equivalent dimensions at 512 x 512. To prevent distortion, images that were smaller than this shape were firstly padded with zeros to have an equivalent row and height ratio (e.g. an image at 324 x 350 was padded to 350 x 350; an image at 375 x 280 was padded to 375 x 375). Images were then upscaled to 512 x 512 using linear interpolation.

Padding before upscaling was done to prevent distortion of the image by maintaining the source image aspect ratio. Upscaling is preferred to downscaling to prevent loss of potentially important information that could be leveraged by the network [Tan *et al.* 2021]. Denoising, registration and windowing were not undertaken: images appear to have adequate registration; all images have been processed by the LONI Quality Control System (LONI QC) before publication which secures consistent data quality through, "range checking, signal-to-noise ratio checks, artifact removal techniques, power spectrum analysis, and various filtering methods" [Khan *et al.* 2022]. Figure 3.4 is an image from Site 2, pre and post processing.

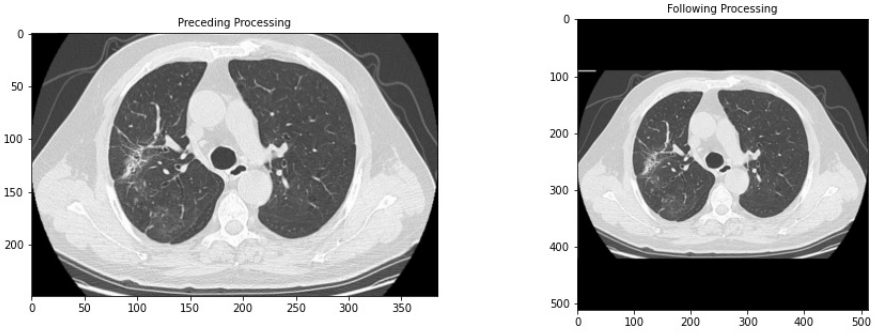


Figure 3.4: Sample of Images Before and After Processing

3.2.2 Model

The research leverages a customized VGG16. The model's dense layers were dropped and replaced with two fully-connected dense layers with shapes 256 and 128 before a two node output. Base layers are pre-trained from ImageNet, with a 512×512 input layer. Table 3.3 illustrates the model's complete architecture where there are 48,302,530 trainable parameters. This is significantly smaller than the standard VGG16 architecture which has nearly 144,000,000 trainable parameters.

The customised model was motivated by similar studies where customised VGG16's with fewer or narrower dense layers achieved comparable levels of accuracy to standard VGG16 networks. It was also motivated by the need to prevent out-of-memory errors that arose due to expansion in the number of parameters due to a larger input layer, and the need to operate concurrent networks for the federated model. Finally, it was motivated by the desire for having more efficient training given the extensive

Table 3.3: Customised VGG16 Network Architecture

Layer (type)	Output Shape	Param #
input_2 (InputLayer)	[(None, 512, 512, 3)]	0
block1_conv1 (Conv2D)	(None, 512, 512, 64)	1,792
block1_conv2 (Conv2D)	(None, 512, 512, 64)	36,928
block1_pool (MaxPooling2D)	(None, 256, 256, 64)	0
block2_conv1 (Conv2D)	(None, 256, 256, 128)	73,856
block2_conv2 (Conv2D)	(None, 256, 256, 128)	147584
block2_pool (MaxPooling2D)	(None, 128, 128, 128)	0
block3_conv1 (Conv2D)	(None, 128, 128, 256)	295,168
block3_conv2 (Conv2D)	(None, 128, 128, 256)	590,080
block3_conv3 (Conv2D)	(None, 128, 128, 256)	590,080
block3_pool (MaxPooling2D)	(None, 64, 64, 256)	0
block4_conv1 (Conv2D)	(None, 64, 64, 512)	1,180,160
block4_conv2 (Conv2D)	(None, 64, 64, 512)	2,359,808
block4_conv3 (Conv2D)	(None, 64, 64, 512)	2,359,808
block4_pool (MaxPooling2D)	(None, 32, 32, 512)	0
block5_conv1 (Conv2D)	(None, 32, 32, 512)	2,359,808
block5_conv2 (Conv2D)	(None, 32, 32, 512)	2,359,808
block5_conv3 (Conv2D)	(None, 32, 32, 512)	2,359,808
block5_pool (MaxPooling2D)	(None, 16, 16, 512)	0
flatten (Flatten)	(None, 131072)	0
dense (Dense)	(None, 256)	33,554,688
dense_1 (Dense)	(None, 128)	32,896
dense_2 (Dense)	(None, 2)	258

Total params: 48,302,530
Trainable params: 48,302,530

experimental testing regime which is discussed below. A comprehensive battery of tests with different combinations of dense layers was evaluated and highlights the desirability of the proposed model. Each combination was evaluated using Site 10_1. The results of these tests are contained in Table 3.4, where there is no significant or systematic difference in test accuracy created by variations in dense layer shape. However, there are improvements in training time of up to 25% result from combinations of dense layers that reduce the number of trainable parameters. All models were evaluated using a Tesla V100-SXM2-16GB GPU.

A softmax activation function is used with a categorical cross-entropy loss. No dropout is used and convolutions are passed through a ReLU layer. The model and associated hyper-parameters were elected through experimentation as above, and upon review of literature.

3.2.3 Testing Regime and Hyperparameters

The research evaluated and compared three sets of models: independent models, global models and federated models. Comparing the results of these models highlight the

Table 3.4: Model Architecture Evaluation

Dense 1	Dense 2	Parameters	Test Accuracy	Training Time
4096	4096	568,379,202	0.966	00: 14: 45
4096	2048	559,984,450	0.984	00: 14: 31
4096	1024	555,787,074	0.986	00: 14: 16
4096	512	553,688,386	0.972	00: 14: 37
4096	256	552,639,042	0.964	00: 14: 56
4096	128	552,114,370	0.980	00: 15: 11
2048	2048	287,352,642	0.988	00: 14: 52
2048	1024	285,252,418	0.986	00: 14: 41
2048	512	284,202,306	0.982	00: 13: 04
2048	256	283,677,250	0.972	00: 13: 02
2048	128	283,414,722	0.982	00: 12: 17
1024	1024	149,985,090	0.974	00: 13: 11
1024	512	149,459,266	0.978	00: 12: 04
1024	256	149,196,354	0.982	00: 11: 43
1024	128	149,064,898	0.972	00: 11: 46
512	512	82,087,746	0.968	00: 11: 12
512	256	81,955,906	0.990	00: 11: 16
512	128	81,889,986	0.970	00: 11: 19
256	256	48,335,682	0.992	00: 11: 00
256	128	48,302,530	0.986	00: 11: 04

feasibility of using a model to diagnose COVID-19 using data for a site that has not contributed data to training the model.

Independent models refer to the CNNs trained using a single site's data. This set of models replicates a siloed approach to developing COVID-19 diagnostics where each site only has access to their own data for training.

Global models refers to CNNs trained on centrally pooled site level data. This approach provides a 'best-case' scenario where patient data is freely shareable. This is the benchmark against which the federated learning model is evaluated. Sharing of patient data between sites would be required in this scenario.

Federated models refers to the use of a federated averaging approach. This produces an openly-accessible network that is constructed using the weights from local networks. This approach replicates the collective development of a COVID-19 diagnostic solution by a set of sites that cannot or are reluctant to share their data. Sharing of patient data would not be required in this scenario.

A facility decision-maker would prefer the federated model if the model has superior test accuracy for data from their site relative to the test accuracy of their own independent model.

The performance of each of the independent, global and federated models was evaluated through two accuracy measures. Accuracy divides the number of correctly classified images by the total number of images. The research's objective determines accuracy as the sole and preferred metric for evaluation. Decision-makers at facilities choosing whether or not to use the federated model, their own model, or neither would rely on the accuracy measure as they are unable to control for the balance in the data of patients presenting for diagnostics. Furthermore, additional metrics such as area under the curve or confusion matrices provide nuanced insight into the behaviour of specific models but do not contribute significantly to understanding the feasibility of using federated learning to enable diagnostics by lower-resourced sites. Optimal weights were stored for the highest accuracy on the validation dataset. These weights were used when evaluating test accuracy. Accuracy was evaluated in two ways:

1. **Internal test accuracy:** the model's accuracy was evaluated using a test set of data from the same site(s) used in training. This highlights generalisation accuracy for the sites used in training.

2. **External test accuracy:** the model's accuracy was evaluated using the test dataset from all of the sites not used in training. This highlights generalisation accuracy of the model across all sites.

A learning rate of 0.00005 enables training of all local models, and was also used in training of the federated and global models. The learning rate was discovered through experimentation, was not decayed, and early stopping was not undertaken to enable observation of training behaviour and tendency for over-fitting. The local and global models were trained for 150 epochs, and the federated models trained for 150 communication rounds. Each model was trained with a batch size of 10, with 20 steps in each training epoch and 10 steps in validation. While this does not guarantee that the

Table 3.5: Training Regime Overview

Regime	Sites Included	Number of Samples
Regime 1	Site 10_1 & Site 10_2	34,310
Regime 2	Site 10_1 & Site 10_2 & Site 7	44,271
Regime 3	Site 10_1 & Site 10_2 & Site 7 & Site 1	47,197

network is exposed to all training samples in each epoch, this is not considered a risk due to the scale of the training data.

Kandel and Castelli [2020] highlighted the impact of batch size and learning rate on the performance of CNNs for diagnostics on a histopathology dataset. By varying the batch size and learning rate of a VGG16 network, the authors illustrate that smaller batch sizes with lower learning rates improve accuracy, suggesting the small batch size and learning rate elected in this research may strengthen results. This finding motivated the decision to use a smaller batch sizes, however, the effect of varying batch size on the accuracy of this research is untested.

Three global and federated models were trained: the first model used two sites, the second used three sites, the third used four sites. This regime provides for an investigation into the generalisation impact of introducing more sites. Table 3.5 highlights the regime, where sites were incrementally added based on the sample size. For the federated model, weights were stored based on accuracy for the federated validation dataset.

All sites were decomposed into training, validation and test sets. These datasets are fixed and disjoint (i.e. images were permanently allocated to training, validation and test, and no image exists in more than on set). Independent, global and federated models were all trained using the same datasets. Training, validation and test data was input using the Keras flow-from-directory module. This was undertaken to limit the burden on memory where Site 10_1's training data, for example, is 5.03 gigabytes. The images were split randomly, with 75% allocated to training, 15% allocated to validation and 10% allocated to test.

Models were evaluated on 50 images randomly selected from the test datasets. All training, validation and test images were shuffled. As is highlighted below, the principle of 'standardised models, data and testing regimes' was adopted to ensure comparability.

Global and Federated Regime Details

Global models were equivalent in architecture and training regime as the independent models. The global models training, validation and internal test sets were constructed using composite data generators. Each batch consisted of an equal number of samples from each site, instead of including a weighted number of samples based on the size of the site's dataset. This was motivated by the objective of maximising external test accuracy - exposing the model to a wider variety of between-site variation instead of within-site variation. Evaluating this hypothesis is beyond the scope of this research.

Federated models were equivalent in architecture to independent and global models. During each communication round, each local site trains for a single epoch, with the same number of steps and batch sizes equivalent to the independent and global models. Following an epoch for each of the local sites, the model weights are averaged and used to update the federated model. The federated model is then evaluated on the federated model validation data. The federated model weights are then used to update the local models for the next communication round. The averaging does not 'weight' the model weights by the number of samples from the site. This is motivated by the same logic detailed in the global model where the research aims to maximise between site variation. Furthermore, each local site has an equivalent number of samples used in training in each communication round.

Chapter 4

Results

The training behaviour and training, validation and test accuracy of each of the independent, global and federated models is discussed in this chapter. This includes internal and external test accuracy. The results indicate that generalisation to Site 2 is underwhelming for all regimes. The results do not appear to warrant the use of a federated model as an alternative to an independent model, or as a complement or alternative to radiographers who are noted in one study as achieving a 70% to 75% accuracy [Nicola Flor and Brambilla 2021].

4.1 Independent Models

Table 4.1 is a summary of the test accuracies of the independent models when evaluated using test data from each of the sites. For example, the first column highlights the test accuracy of the independent model trained using Site 1 data. The rows indicate the test accuracy of this model when evaluated on test data across the different sites. This indicates that Site 1 achieved a 0.976 accuracy with Site 1 test data. The table also indicates external test accuracy where Site 1 generalises best to Site 10_2 at 0.698 and worst to Site 7 at 0.298. A decision-maker would compare the accuracy of their internal test accuracy to the accuracy of a federated model to choose which model to use.

The results can be averaged along the columns to reveal each model's average external generalisation capacity. This is indicated in Table 4.2 which contains the average test accuracy of the model when evaluated on test data from each of the sites excluding itself. This highlights that Site 10_2 is most able to generalise to external sites, and Site 10_1 the least able to.

The results can be averaged along the rows to reveal how challenging different sites are to generalise towards. This is indicated in Table 4.3 which contains the average test accuracy of the site when evaluated with the models that were not trained with that site's data. For example, Site 1 appears to be the most difficult site to generalise to by averaging the test accuracies for Site 1 test data when using models trained on Site 2,

Table 4.1: Independent Model Test Accuracy Matrix

	Site 1 (Trained)	Site 2 (Trained)	Site 7 (Trained)	Site10_1 (Trained)	Site10_2 (Trained)
Site 1 (Eval.)	97.6%	49.1%	62.8%	25.6%	71.3%
Site 2 (Eval.)	49.3%	89.3%	49.3%	54.6%	62.7%
Site 7 (Eval.)	29.8%	77.4%	95.6%	75.8%	69.4%
Site10_1 (Eval.)	60.6%	52.2%	63.2%	98.4%	63.8%
Site10_2 (Eval.)	69.8%	65.0%	52.2%	34.6%	93.6%

Note: The cells highlighted in **green** contain the accuracy of a model that is evaluated using test data from the same site that was used in training the model.

Table 4.2: Independent Model Average External Test Accuracy

	Site1 (Trained)	Site 2 (Trained)	Site 7 (Trained)	Site10_1 (Trained)	Site10_2 (Trained)
Avg Accuracy (External)	52.4%	60.9%	56.9%	47.7%	66.8%

Table 4.3: Average Test Accuracy per Site for Models Trained on Other Site Data.

	Site 1 (Eval.)	Site 2 (Eval.)	Site 7 (Eval.)	Site10_1 (Eval.)	Site10_2 (Eval.)
Avg Accuracy (Generalise to)	52.2%	54.0%	63.1%	60.0%	55.4%

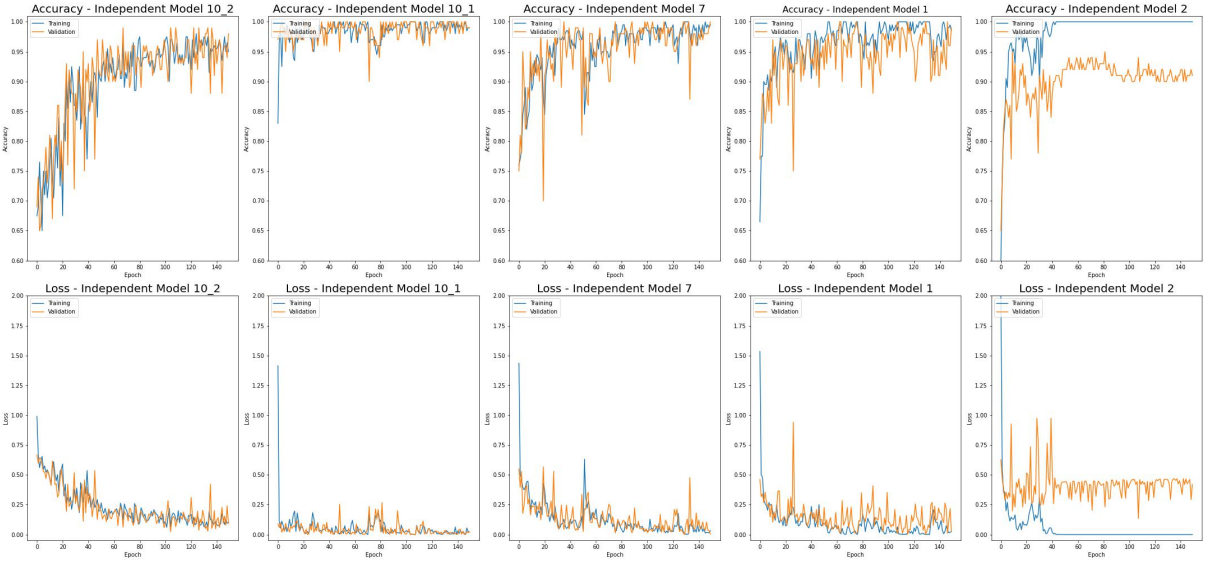


Figure 4.1: Independent Models Training and Loss

Site 7, Site 10_1 and Site 10_2. In contrast, Site 10_1 appears the easiest to generalise to.

Figure 4.1 provides an overview of the training and validation accuracy and loss of the models during training. The following subsections explore these dynamics.

4.1.1 Independent Model: Site 10_2

Figure 4.1 highlights how the site achieved a maximum training accuracy of 0.98, a maximum validation accuracy of 0.99 and a testing accuracy of 0.93. The maximum validation accuracy was first reached during epoch 67: the network converged slowly relative to other independent models. The site was stable in training relative to other sites with no observed collapses in accuracy. The site generalised best to Site 1 and poorest to Site 2, with an average external test accuracy of 0.66.

4.1.2 Independent Model: Site 10_1

The site achieved a maximum training accuracy of 1.0, a maximum validation accuracy of 1.0 and a testing accuracy of 0.98. The maximum validation accuracy was first reached during epoch 3: the network converged very quickly relative to other indepen-

Table 4.4: Global Model Internal Test Accuracy

	Global_1	Global_2	Global_3
Test Accuracy (Internal)	94.2%	97.2%	96.1%

dent models. The site was stable in training relative to other site training behaviour with a single but quickly recovered collapse in accuracy. The site generalised best to Site 7 and poorest to Site 1, with an average external test accuracy of 0.47.

4.1.3 Independent Model: Site 7

Figure 4.1 highlights how the site achieved a maximum training accuracy of 1.0, a maximum validation accuracy of 1.0 and a testing accuracy of 0.95. The maximum validation accuracy was first reached during epoch 36: the network converged moderately quickly relative to other independent models. The site was less stable in training relative to other site training behaviour with few instances of collapse in accuracy. The site generalised best to Site 10_1 and poorest to Site 2, with an average external test accuracy of 0.56.

4.1.4 Independent Model: Site 2

Figure 4.1 highlights how the site achieved a maximum training accuracy of 1.0, a maximum validation accuracy of 0.95 and a testing accuracy of 0.89. The maximum validation accuracy was first reached during epoch 9: the network converged quickly relative to other independent models. The site was stable in training relative to other sites. The site generalised best to Site 10_2 and poorest to Site 7, with an average external test accuracy 0.61.

4.1.5 Independent Model: Site 1

Figure 4.1 highlights how the site achieved a maximum training accuracy of 1.0, a maximum validation accuracy of 1.0 and a testing accuracy of 0.97. The maximum validation accuracy was first reached during epoch 62: the network converged slowly relative to other independent models. The site generalised best to Site 7 and poorest to Site 1, with an average external test accuracy of 0.52.

4.2 Global Models

All of the global models achieved competitive test accuracies when evaluated using the global model's own test data. This reveals that each global converged and provides accurate diagnostics for images from the sites contributing training data to the global model. The test accuracies of the global models are highlighted in Table 4.4. This indicates internal test accuracy.

Table 4.5: Global Model Site Specific Internal and External Test Accuracy

	Global_1	Global_2	Global_3
Site 2 (Evaluated)	54.7%	50.6%	60.0%
Site 1 (Evaluated)	29.7%	72.7%	94.5%
Site 7 (Evaluated)	77.8%	98.2%	98.0%
Site10_1 (Evaluated)	99.2%	98.7%	99.2%
Site10_2 (Evaluated)	92.0%	94.3%	91.5%

Note: The cells highlighted in **green** are sites included in training the model. Cells in **red** are sites that were not included in training.

Table 4.6: Global Model Average External Test Accuracy

	Global_1	Global_2	Global_3
Test Accuracy (Average External)	54.1%	61.7%	60.0%

The global models were also evaluated using the test data from each of the sites. This details the accuracy of the global model in diagnosing images from sites that contributed data in training, and sites that did not contribute training data. This indicates site specific internal and external test accuracies. This is summarised in Table 4.5 where the columns indicate the model and the rows indicate the sites used in testing. The cells highlighted in green contributed data for training for the associated global model, and in red did not contribute data for training. For example, Global Model 1, was trained using data from Site 10_1 and 10_2. Green cells therefore indicate internal test accuracy and red cells external test accuracy. This indicates that internal test accuracies across the global models were at least 0.92. It also indicates that no global model could achieve more than 0.60 test accuracy for Site 2 and that Global Model 1 was able to achieve 0.78 accuracy for Site 7.

Each global model's capacity to generalise to other sites is revealed by averaging the test accuracies of the sites not used in training (i.e. average external test accuracy). These averages are contained in Table 4.6. This indicates that none of the global models achieve useful test accuracies for sites not included in training.

Figure 4.2 provides an overview of the training and validation accuracy and loss of the models during training. The following subsections explore these dynamics.

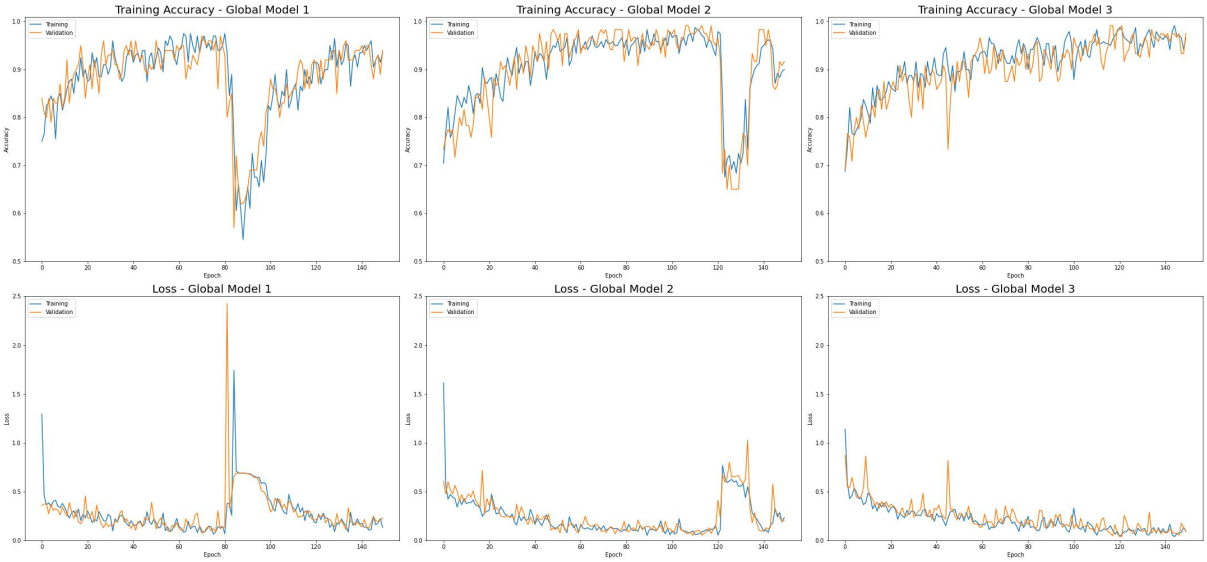


Figure 4.2: Global Models Accuracy and Loss

4.2.1 Global Model 1: Site 10_1 and 10_2

Figure 4.2 highlights how the model achieved a maximum training accuracy of 0.98, a maximum validation accuracy of 1.00 and a testing accuracy of 0.942. The maximum validation accuracy was first reached during epoch 141: the network converged moderately quickly relative to other global models. The global model was stable in training however it collapsed between epochs 80 and 90. Of the external sites, the model generalised best to Site 7 and poorest to Site 1, with an average external test accuracy of 0.541.

4.2.2 Global Model 2: Site 10_1 & Site 10_2 & Site 7

Figure 4.2 highlights how the model achieved a maximum training accuracy of 0.98, a maximum validation accuracy of 1.00 and a testing accuracy of 0.97. The maximum validation accuracy was first reached during epoch 119: the network converged moderately relative to other global models. The global model was stable in training however it collapsed between epochs 120 and 130. Of the external sites, the model generalised best to Site 1 and poorest to Site 2, with an average external test accuracy of 0.62.

4.2.3 Global Model 3: Site 10_1 & Site 10_2 & Site 7 & Site 1

Figure 4.2 highlights how the model achieved a maximum training accuracy of 0.99, a maximum validation accuracy of 0.98 and a testing accuracy of 0.96. The maximum validation accuracy was first reached during epoch 147: the network converged slowly relative to other global models. The global model was less stable in training however it did not collapse. The model's external test accuracy was evaluated against Site 2, achieving 0.60.

Table 4.7: Federated Model Internal Test Accuracy

	Federated_1	Federated_2	Federated_3
Test Accuracy (Internal)	98.0%	99.0%	98.5%

Table 4.8: Federated Model Site Specific Internal and External Test Accuracy

	Federated_1	Federated_2	Federated_3
Site 2 (Evaluated)	52.0%	52.0%	57.3%
Site 1 (Evaluated)	32.1%	68.6%	97.3%
Site 7 (Evaluated)	67.0%	98.8%	99.5%
Site10_1 (Evaluated)	99.2%	99.2%	99.7%
Site10_2 (Evaluated)	97.2%	97.5%	96.3%

Note: The cells highlighted in **green** are sites included in training the model. Cells in **red** are sites that were not included in training.

4.3 Federated Models

All of the federated models also achieved competitive test accuracies when evaluated using the federated model's own test data. This reveals that each federated model converged and provides accurate diagnostics for images from the sites contributing training data to the federated model. The test accuracies of the federated models are highlighted in Table 4.7. This indicates internal test accuracy. The federated models achieved comparable results to the global models. The federated models were also evaluated using the test data from each of the sites. This details the accuracy of the federated model in diagnosing images from sites that contributed data in training, and sites that did not contribute training data. This indicates site specific internal and external test accuracies. This is summarised in Table 4.8 where the columns indicate the model and the rows indicate the sites used in testing. The cells highlighted in green contributed data in training the associated federated model, and in red did not contributed data in training. This indicates that internal test accuracies across the federated models were at least 0.96. It also indicates that no global model could achieve more than 0.57 test accuracy for Site 2 and that Federated Model 1 was able to achieve 0.77 accuracy for Site 7. Each federated model's capacity to generalise to other sites is revealed by averaging the test accuracies of the sites not used in training (i.e. average external test accuracy). These averages are contained in Table 4.9. This indicates that none of the federated models achieve useful test accuracies for sites not included in training. Figure 4.3 provides an overview of the test accuracy and loss of the models during training, and a decomposed view of training behavior of each of sites used as local models. The following subsections explore these dynamics.

Table 4.9: Federated Model Average External Test Accuracy

	Federated_1	Federated_2	Federated_3
Test Accuracy (Average External)	50.37%	60.3%	57.3%

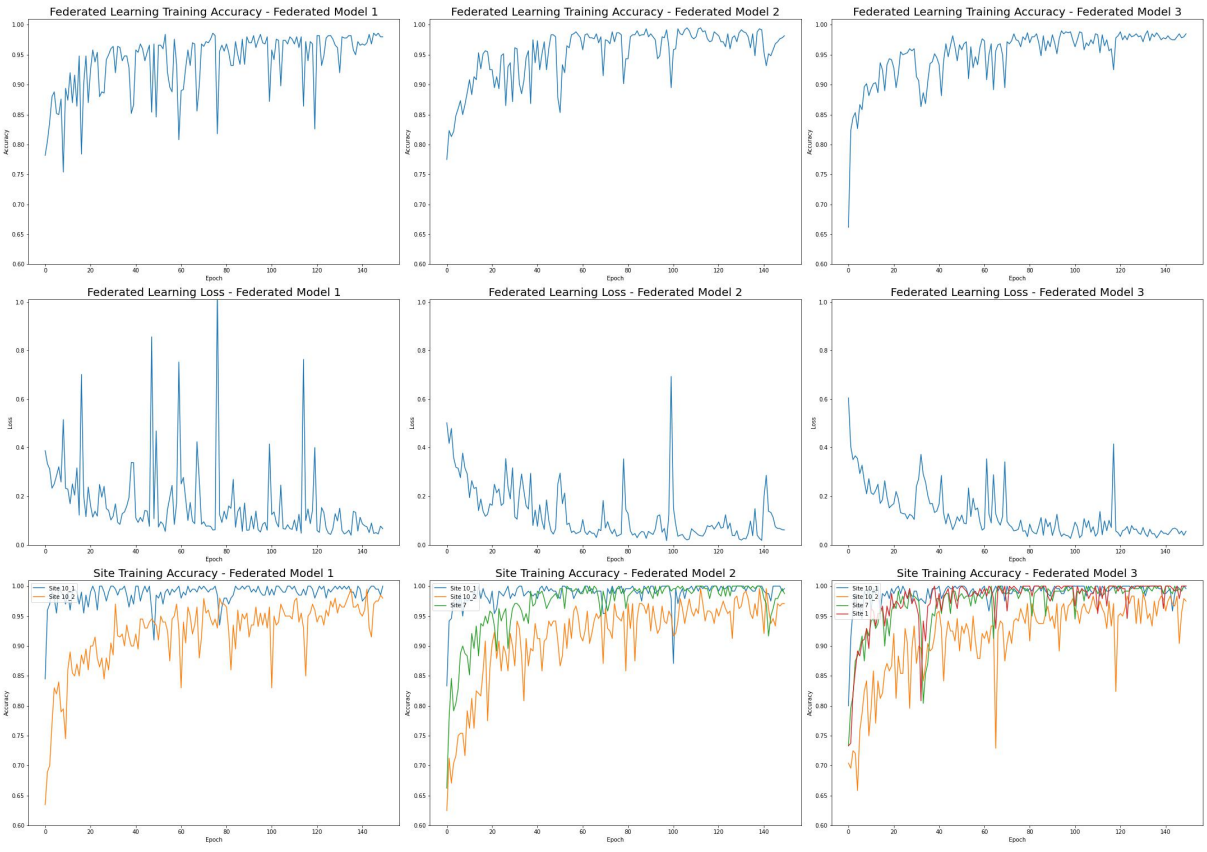


Figure 4.3: Federated Learning Models Accuracy and Loss

4.3.1 Federated Model 1: Site 10_1 and 10_2

The model achieved a maximum validation accuracy of 0.986 and test accuracy of 0.980. The maximum validation accuracy was first reached during epoch 74: the network converged quickly relative to other federated models. The model was moderately stable with a few minor instances of collapse. Of the external sites, the model generalised best to Site 7 and poorest to Site 1, with an average external test accuracy of 0.504.

4.3.2 Federated Model 2: Site 10_1 & Site 10_2 & Site 7

The model achieved a maximum validation accuracy of 0.995 and test accuracy of 0.990. The maximum validation accuracy was first reached during epoch 106: the network converged moderately quickly relative to other federated models. The model was moderately stable with a few minor instances of collapse. Of the external sites, the model generalised best to Site 1 and poorest to Site 2, with an average external test accuracy of 0.603.

4.3.3 Federated Model 3: Site 10_1 & Site 10_2 & Site 7 & Site 1

The model achieved a maximum validation accuracy of 0.990 and test accuracy of 0.985. The maximum validation accuracy was first reached during epoch 94: the network converged moderately quickly relative to other federated models. The model was moderately stable with a few minor instances of collapse. The site achieved an external test accuracy of 0.573 for Site 2.

Chapter 5

Discussion

The results provide important insights into the feasibility of leveraging federated learning to enable diagnostics of CT scans by lower-resourced sites that may not be positioned to contribute training data, or may have the option of developing their own independent model. The poor external test accuracy to Site 2 highlights that the elected model, pre-processing approach and training regimes do not result in a federated model with accuracies superior to independent models, or to radiography experts. As such, a healthcare facility decision-maker is unlikely to prefer the federated model to these other options. This result indicates the need for further research to evaluate different conditions that might improve these results. A discussion of the results and areas of further research are contained in this chapter.

5.1 Results

The performance and behaviours of the independent models provide important insights into the prerequisites for external generalisation. While most models achieved internal test accuracies comparable with the literature, none managed to achieve a useful level of external test accuracy, nor an external test accuracy that is superior to the test accuracies of the independent models of the sites not included in training.

Three observations provide insight: firstly, Site 2 achieved the poorest internal test accuracy suggesting that it is challenging to achieve high levels of accuracy when training using data from a variety of sites; secondly, there is little correlation between average external test accuracy and the number of samples used in training, suggesting that the higher volumes of training data do not necessarily result in high external test accuracy; finally, Site 10_2 (which had non-lung areas masked during pre-processing) achieved the greatest external test accuracy, suggesting that this step (and potentially others) in pre-processing may be key in improving external test accuracy.

The global models achieved internal test accuracies comparable with the literature and the independent models at 0.942 to 0.972. Separately evaluating the accuracy of the global models on each of the sites' test data indicates similar accuracies to the internal test accuracies of the global models. For example, Global Model 1 achieves 0.920 and

0.992 accuracy when evaluated on Site 10_2 and Site 10_1 respectively; the internal test accuracy of the independent models of these sites are 0.936 and 0.984 respectively.

The global models converged more slowly than the independent models with a greater number of epochs needed to converge when more sites are included in training. The global models were furthermore less stable in training than the independent models and were subject to more erratic dips and recoveries in accuracy. This is attributed to the wider variation of data included in training.

As with the independent models, none of the global models achieved competitive external test accuracies when evaluated on Site 2, or any other site not used in training. This suggests that it is not feasible to achieve high levels of external test accuracy using the current model and site data. There were furthermore little systematic increases in external test accuracies when introducing more sites in training (i.e. from Global 1 to Global 2 to Global 3). This suggests either that achieving higher levels of external test accuracy may not necessarily be improved by introducing more sites, or that external test accuracy could be improved by introducing a substantial number of additional sites.

The federated models achieved internal test accuracies superior to their global counterparts. This includes site specific internal test accuracies. This is an unexpected result as the internal test accuracies of the global models were hypothesised as the theoretical upper bound of internal test accuracy for the federated models. The superior results could be attributed to the greater number of training samples used and steps undertaken in each federated model communication round relative to each global model epoch. In terms of training steps for example, each global model epoch undertakes 20 training steps whereas each federated model communication round undertakes 20 training steps for each of the local models. In terms of training samples for example, each global model epoch exposes the network to training samples equal in number to the steps x the batch size. In contrast, each federated model communication round results in each of the local networks being exposed to training samples equal in number to the steps x the batch size. This may suggest that increases in batch size across all models may improve results.

The federated models achieved comparable external test accuracy as their global counterparts and likewise were unsuccessful in achieving a practically useful external test accuracy for Site 2. The federated models' site specific internal and external test accuracies were similar to their global counterparts. For example, Global Model 1 achieved an external test accuracy of 0.778 for Site 7, whereas Federated Model 1 achieved 0.670. The federated models likewise do not indicate that introducing additional sites improves external test accuracy. These results indicate that a healthcare facility decision maker would not opt-in to the federated model as an alternative to their own independent model, and are unlikely to use the federated model as a complement or alternative to their own radiographers.

The federated models converged at a speed comparable to their global counterparts. In contrast, the federated models were observed to be less stable and were more inclined to intermittent, minor dips in accuracy however these are recovered quickly. These

intermittent dips are suspected to be instances where a local model achieves a rapid improvement in accuracy during a communication round that results in large weight updates which dominate the federated averaging update and therefore deteriorates generalisation accuracy of the federated model (i.e. the averaged weights are more oriented towards predictions for data from the dominant local model than data from the complete validation set). Inspection of the validation accuracy of Federated Model 1 and the accuracy of its associated sites does not offer evidence to affirm this hypothesis, however indicates that a collapse in the federated model's accuracy is often followed by a collapse in accuracy of one or both of the local models. Further analysis of this behaviour is beyond the scope of this research.

The inability of the networks to achieve high levels of external test accuracy for Site 2 is not attributable to overfitting on internal data. The hypothesis that earlier stopping could result in improved performance was explored: Global Model 3 was trained again, however the validation data was replaced with the complete Site 2 dataset. Figure 5.1 highlights that test accuracy for Site 2 ranged between 0.4 and 0.7 throughout training.

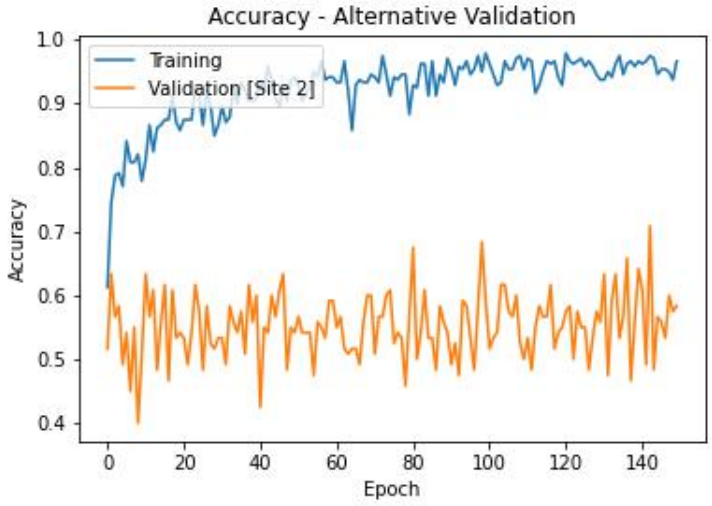


Figure 5.1: Global Model 3 Training and Validation Accuracy

5.2 Further Research

The results indicate that collaborative development of diagnostic models to enable their use by sites non-participant in training requires further research. This research could be extended in six ways, highlighted below. Areas of further research reflect the limitations of the study.

Firstly, exploring alternative network architectures. The customised VGG16 used in this research could yield additional generalisation capability through expansion in the size of dense layers. However, the expected gains in external test accuracy are expected to be negligible given that the current network is able to achieve high levels of internal test accuracy. Preliminary testing furthermore highlights that varying the shapes of the

dense layers result in negligible changes in test accuracy. It is also expected that the introduction of dropout layers would have a negligible impact on external test accuracy - Figure 5.1 indicates that the poor external test accuracy is unlikely a function of overfitting.

Secondly, weighting the number of samples from each site used in training the global and federated models according to the number of training samples available for that site. This would increase the network's exposure to data from sites with larger training datasets. This is expected to have a negligible impact on external test accuracy as the global and federated models achieved high levels of internal test accuracy. Furthermore, all independent, federated and global models achieved poor external test accuracy.

Thirdly, exploring the impact of customised hyper-parameter tuning for each of the local models in the federated training regime. Zawad *et al.* [2021] conduct a series of experiments which highlight that the performance of a federated learning model can be strengthened by designing hyper-parameters for each client that are reflective of the underlying data distribution and quantity of each client. The authors' proposed FedTune methodology is not expected to improve external test accuracy significantly more than the previous item [Zawad *et al.* 2021].

Fourthly, introducing 'weighting' to the federated averaging step, where the local network weights during averaging are weighted according to the number of training samples available to that local site. This is likewise not expected to improve external generalization capacity.

Fifthly, exploring the implications of alternative approaches to data preprocessing. This includes the steps proposed by Masoudi *et al.* [2021] highlighted in Section 5 or an approach similar to Heidari *et al.* [2020] highlighted in Section 4. This would derive more comparable inputs across sites. The leading external test accuracy of Site 10_2 - where non-lung areas have been masked - suggests that this could yield meaningful improvements in external test accuracy. Data augmentation to increase the number of samples for sites with fewer samples could also be explored. Unless this data augmentation is undertaken in a way that increases variation (and potentially noise), it is hypothesised that this would not meaningfully improve external test accuracy, as all models are already converging to commendable internal test accuracies.

Finally, exploring sensitivities in external test accuracy when including a small number of composite site images in training, such as Site 2. This would expose the network to samples with wide variation. This is hypothesised to result in the greatest improvements in external generalisation.

Chapter 6

Conclusion

Federated learning enables the development of diagnostic solutions without having to centralise data used in training. The COVID-19 pandemic stimulated the release of chest CT scans that can be used in research into AI-enabled diagnostics. Segmentation of these scans by sites enables their use in federated learning regimes. The research demonstrates that a customised VGG-16 in a federated learning regime achieves test accuracies comparable to models trained with centralised data when evaluated on data from sites used in training. However, models trained in a federated learning regime or with centralised data is unable to generalise to data from sites not used in training. Achieving this would offer significant value for the health-sector by providing lower-resourced facilities with a tool to support radiographers, or an alternative to developing their own independent models, however the findings suggest further research should be undertaken to evaluate whether and how image preprocessing can improve results.

Appendix A

Supplementary Model Details

Classification	Title [Authors]	Overview
Federated Learning for COVID-19	COVID-19 detection using federated machine learning [Salam et. al.]	**Objective:** The authors aim to compare the accuracy of FL and traditional machine learning approaches to diagnosing COVID-19 through chest x-rays. **Model Description(s):** The authors compared the performance of federated averaging and traditional machine learning. Both used a sequential deep learning model. The authors experimented with SGD and ADAM optimizers, and binary cross-entropy and sparse categorical cross-entropy loss functions. The authors also varied the learning rates. **Findings and Learnings:** The FL model with SGD algorithm had a higher prediction accuracy, higher training time and a lower prediction loss than the traditional model. The softmax activation function and SGD optimizer provided better prediction accuracy and loss. **Limitations:** The authors do not make use of a CNN for the classification of images. The authors do not provide clarity on how the images are segmented in the FL model. The input data that the authors refer to does not contain meta-data that would enable site-level segmentation to meaningfully replicate real-world conditions for collective development of a diagnostic solution. It is assumed that the images were randomly segmented

Federated Learning and CNN for COVID-19	Experiments of Federated Learning for COVID-19 Chest X-ray Images [Liu et. al.]	**Objective:** The authors aim to compare the accuracy of FL and traditional machine learning approaches to diagnosing COVID-19 through chest x-rays. The authors also compare the performance of four models in the FL and traditional machine learning approaches **Model Description(s):** The authors compared the performance of federated averaging and traditional machine learning. The authors compare four networks: CovidNet, ResNeXt, MobileNet-v2, and ResNet18. **Findings and Learnings:** All networks in the FL model converged slower and to a higher loss than without FL. The authors also found that ResNet18 has the fastest convergence speed and highest accuracy. **Limitations:** The authors do not provide clarity on how the images are segmented in the FL model. The authors note that they define a set of image owners, allocate images to the owners and assume that the owners are different hospitals. This appears similar to randomly splitting images into different 'owners'.
Machine Learning for COVID-19	Collaborative City Digital Twin For Covid-19 Pandemic: A Federated Learning Solution [Pang et. al.]	**Objective:** The authors aim to demonstrate that a FL model can be used to predict the impact of city-level policy interventions using historical infection and intervention event data from different cities. **Model Description(s):** The authors compared the performance of federated averaging and traditional machine learning. The authors use a Seq2seq Encoder-Decoder network. The authors segment the data by state. Data contains daily infections and historical interventions. **Findings and Learnings:** The authors find that the FL model provides superior accuracy. The authors also find that the model's accuracy increases as the number of participants is increased. **Limitations:** The model demonstrates the opportunity for FL in the optimisation of COVID-19 policy response but does not provide for COVID-19 diagnostics.

CNN for COVID-19	Insight about detection, prediction and weather impact of COVID-19 using neural network [Haque et. al]	**Objective:** The authors aim to demonstrate differences in the accuracy of classifying COVID-19 chest X-Rays using a range of CNNs. The authors also explored the impact of weather factors on the rate of infection and likelihood of death. **Model Description(s):** The authors compared five pre-trained CNNs. This includes VGG16, VGG19, Xception, InceptionV3 and Resnet50. The models made use of a SoftMax activation function to classify images into 1 of 3 groups. The authors used a categorical cross-entropy loss function and ADAM optimizer. **Findings and Learnings:** The VGG16 and VGG19 models demonstrated the highest accuracy after training. Both achieved over 92% accuracy for the validation set. **Limitations:** The study sample was small at 200 images. Furthermore, there was no segmentation of the data nor use of FL. The paper therefore does not adequately reflect the complexities of developing diagnostic solutions while preserving the privacy of the underlying data.
Machine Learning for COVID-19	The Number of Confirmed Cases of Covid-19 by using Machine Learning: Methods and Challenge [Ahmad et. al]]	**Objective:** the paper provides a detailed literature review of machine learning approaches to the prediction of COVID-19 cases to help machine learning practitioners develop more accurate and robust models. **Model Description(s):** Not Applicable **Findings and Learnings:** The authors present a taxonomy of models used for the prediction of COVID-19. There are four groups. **Limitations:** The paper does not facilitate research into COVID-19 diagnostics.

Federated Learning and CNN for COVID-19	FedNAS: Federated Deep Learning via Neural Architecture Search [Chaoyang et. al]	**Objective:** the authors aim to demonstrate the effectiveness of FL in diagnosing COVID-19 using CT images. The authors illustrate that the non-IID nature of the data creates complexities for determining an appropriate deep learning architecture. **Model Description(s):** The authors present a novel AutoFL model – FedNAS (Federate Neural Architecture Search) – which improves the accuracy of FL by searching for architectures that deliver better results. The authors vary the hyperparameters and the CNN network architecture. **Findings and Learnings:** the study illustrates that FedNAS yields higher accuracy in the classification of the CIFAR10 dataset than FedAvg (81.2% vs 77.8%). **Limitations:** the authors highlight that the findings will be useful in diagnostics for COVID-19 using CT images however do not evaluate their model with this use case.
CNN for COVID-19	Multi Deep Learning to Diagnose COVID-19 in Lung X-Ray Images with Majority Vote Technique [Fibriani et. al]	**Objective:** the authors aim to demonstrate that a majority voting approach to classifying COVID-19 using a set of CNNs yields superior results. The authors aim to highlight that the majority voting approach minimizes errors relative to a single CNN. **Model Description(s):** the authors create a system of nine different CNNs which are combined using three different majority voting techniques. The CNNs the authors tested includes: AlexNet, ResNet 50, Wide ResNet 50 – 2, ResNet 101, ResNet 152, VGG16, VGG16 Batch, Normalization (BN), VGG19, and VGG19 Batch Normalization (BN). **Findings and Learnings:** the 'Soft Majority Vote' model achieved the highest accuracy at 99.3%. The soft majority vote model made use of a weighted majority voting technique. **Limitations:** the research did not segment the data nor use FL. The paper therefore does not adequately reflect the complexities of developing diagnostic solutions while preserving the privacy of the underlying data.

CNN for COVID-19	Deep neural network to detect COVID-19: one architecture for both CT Scans and Chest X-rays [Mukherjee et al.]	**Objective:** the authors aim to illustrate that CNNs trained with both chest x-rays and CT scans will more accurately classify COVID-19 as the training allows the model to learn more anomalous patterns caused by COVID-19. **Model Description(s):** the authors proposed a nine layer CNN. The network was designed through experimentation where the authors varied the number of layers and values of the hyperparameters. The x-ray and CT scan inputs were 100x100. A total of 672 images were used. **Findings and Learnings:** the proposed model achieves a 96.3% accuracy compared to 71.9% for InceptionV3, 82.9% for MobileNet, and 90.8% for ResNet. **Limitations:** the research did not segment the data nor use FL. The paper therefore does not adequately reflect the complexities of developing diagnostic solutions while preserving the privacy of the underlying data.
CNN for COVID-19	Automated Deep Transfer Learning-Based Approach for Detection of COVID-19 Infection in Chest X-rays [Narayan Das. et. al.]	**Objective:** the authors aim to illustrate the accuracy of a deep transfer learning based approach to diagnosing COVID-19 using chest x-rays. The paper appears to be one of the first efforts to develop rapid deep-learning based diagnostic alternatives to RT-PCR kits. The authors compared their proposed approach to a range of CNN and other machine learning classification techniques. **Model Description(s):** the authors made use of Xception with a Softmax activation function. The Xception model was pre-trained on a larger dataset to allow fine-tuning with the smaller COVID-19 x-ray dataset. **Findings and Learnings:** the proposed model achieved the highest accuracy of 97% relative to the set of competitive alternatives. **Limitations:** the research did not segment the data nor use FL. The paper therefore does not adequately reflect the complexities of developing diagnostic solutions while preserving the privacy of the underlying data.

CNN for COVID-19	Interpretable artificial intelligence framework for COVID-19 screening on chest X-ray [Tsiknakis et. al.]	**Objective:** the authors aim to illustrate the superiority of a transfer learning approach to deep learning for the classification of COVID-19 in x-rays. The transfer learning approach aimed to account for the lack of big and relevant datasets. **Model Description(s):** the authors trained InceptionV3 on the ImageNet dataset. The pre-trained model was then fine-tuned with the COVID-19 training data where the 'backbone' layers were frozen. A total of 522 images across four classes were used with confirmed COVID-19 images accounting for $\sim 25\%$ of the dataset. The authors also included the GradCAM algorithm to support interpretability. **Findings and Learnings:** the authors achieved better performance than comparable studies in the binary classification of COVID-19. The model's performance was inferior to comparable studies in the ternary and quaternary classifications. **Limitations:** the research did not segment the data nor use FL. The paper therefore does not adequately reflect the complexities of developing diagnostic solutions while preserving the privacy of the underlying data.
CNN for COVID-19	COVID-19 Detection using Artificial Intelligence [Salman et al.]	**Objective:** the authors aim to illustrate the accuracy of using CNN for diagnosing COVID-19 using chest x-rays. The paper appears to be one of the first efforts to develop rapid deep-learning based diagnostic alternatives to RT-PCR kits. The authors used transfer learning to cater for the small dataset. **Model Description(s):** the authors make use of a weighted FedAVG and a U-shaped fully convolutional network. The authors also implemented a semi-supervised learning approach to using training data from sites without annotations. The authors made use of 1,704 scans from three countries. The authors also created a control population for evaluation. **Findings and Learnings:** the FL model performs better than each local model. The authors also note that generalization accuracy increases as local models are introduced.

		Limitations: the research provides an important foundation for analysis of FL and CNN for the detection of COVID-19. However, the study has a narrow range of sites. Furthermore, one of the sites does not have annotated data.
CNN for COVID-19	Detection and analysis of COVID-19 in medical images using deep learning techniques [Yang et. al.]	**Objective:** the authors aim to illustrate that customised deep-learning networks for image classification can yield superior results. These customised models are pre-trained on imagenet and fine-tuned using. Customisation is undertaken in a variety of ways, including the shape and number of dense layers in VGG16. **Model Description(s):** the authors evaluate a range of models: VGG16 DenseNet121,ResNet50, and ResNet152. These models are pre-trained using ImageNet. The VGG16 architecture is adapted by altering the structure of the dense layers: four dense layers are used before the final output layer. They have the following shapes: 1024, 512, 256 and 128. **Findings and Learnings:** the customised models achieve competitive accuracies. For example, VGG16 achieves 98% accuracy. The authors also highlight that scaling training data improves performance. **Limitations:** the analysis does not engage with the need for privacy preservation in training. It also does not evaluate models on test data not used in training.
CNN for COVID-19	COV-DLS: Prediction of COVID-19 from X-Rays Using Enhanced Deep Transfer Learning Techniques [Kumar et. al.]	**Objective:** the authors aim to illustrate the modifications to well established convolutional neural networks can yield superior results. This includes altering the shape and number of dense layers, and evaluating the impact of including dropout during training. **Model Description(s):** the authors construct a Modified-VGG16, Modified-VGG19, Modified-ResNet50, and Modified-InceptionV3. All images were standardised to 224x224 resolution. The base models were loaded without the tops and heads, where the authors then introduced a single dense layer of size 64 before classification. The authors use an Adam optimiser and Softmax activation.

		Findings and Learnings: The authors achieve competitive performance with a classification accuracy of 98.61% achieved by Modified-VGG16, 97.22% by Modified-VGG19, 95.13% by Modified-ResNet50, and 99.31% by Modified-InceptionV3. The authors find that these results are superior to the unmodified architectures. **Limitations:** the analysis does not engage with the need for privacy preservation in training. It also does not evaluate models on test data not used in training.
CNN for COVID-19	Improving performance of CNN to predict likelihood of COVID-19 using chest X-ray images with preprocessing algorithms [Heidari et. al.]	**Objective:** the authors aim to demonstrate how preprocessing algorithms can improve the accuracy of CNN models used to classify COVID-19 using chest X-ray imagery. The authors undertake this study to strengthen the network's ability to distinguish between COVID-19 and other respiratory diseases. **Model Description(s):** the authors evaluate a pretrained VGG16 which is fine-tuned to the task. Images used in training are pre-processed with the following steps: remove diaphragms, conduct image noise filtering and histogram equalization. **Findings and Learnings:** the authors compare the accuracy of the VGG16 model with and without preprocessing. The authors find that the image preprocessing regime increases accuracy of the model from 87.6% to 91.0%. **Limitations:** the analysis does not engage with the need for privacy preservation in training. It also does not evaluate models on test data not used in training.

References

[Abdul Salam *et al.* 2021] Mustafa Abdul Salam, Sanaa Taha, and Mohamed Ramadan. Covid-19 detection using federated machine learning. *PLoS One*, 16(6):e0252573, 2021.

[Afshar *et al.* 2020] Parnian Afshar, Shahin Heidarian, Farnoosh Naderkhani, Anastasia Oikonomou, Konstantinos N Plataniotis, and Arash Mohammadi. Covid-caps: A capsule network-based framework for identification of covid-19 cases from x-ray images. *Pattern Recognition Letters*, 138:638–643, 2020.

[Arora *et al.* 2020] N Arora, AK Banerjee, and ML Narasu. *The role of artificial intelligence in tackling COVID-19. Futur Virol 15 (11): 717–724*, 2020.

[Bell 2020] D. Bell. Covid-19. reference article. *Radiopaedia.org. (accessed on 26 Jun 2022)*, 2020.

[Brendan McMahan *et al.* 2016] H Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. *arXiv e-prints*, pages arXiv–1602, 2016.

[Chatzitofis *et al.* 2021] Anargyros Chatzitofis, Pierandrea Cancian, Vasileios Gkitsas, Alessandro Carlucci, Panagiotis Stalidis, Georgios Albanis, Antonis Karakottas, Theodoros Semertzidis, Petros Daras, Caterina Giannitto, et al. Volume-of-interest aware deep neural networks for rapid chest ct-based covid-19 patient risk assessment. *International Journal of Environmental Research and Public Health*, 18(6):2842, 2021.

[Chun-Shuang Guan and Chen 2020] Ru-Ming Xie Zhi-Bin Lv Shuo Yan Zi-Xin Zhang Chun-Shuang Guan, Lian-Gui Wei and Bu-Dong Chen. Ct findings of covid-19 in follow-up: comparison between progression and recovery. *Diagn Interv Radiol. Jul 26(4)*, page 301–307, 2020.

[Das *et al.* 2020] N Narayan Das, Naresh Kumar, Manjit Kaur, Vijay Kumar, and Dilbag Singh. Automated deep transfer learning-based approach for detection of covid-19 infection in chest x-rays. *Irbm*, 2020.

[Deng *et al.* 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. pages 248–255, 2009.

[EU 2022] GDPR EU. What is gdpr, the eu's new data protection law? 2022.

[Ferguson *et al.* 2018] Max K Ferguson, AK Ronay, Yung-Tsun Tina Lee, and Kincho H Law. Detection and segmentation of manufacturing defects with convolutional neural networks and transfer learning. *Smart and sustainable manufacturing systems*, 2, 2018.

[Fibriani *et al.* 2020] Ike Fibriani, Widjonarko Widjnorako, Aris Prasetyo, Angga Mardro Raharjo, and Dasapta Erwin Irawan. Multi deep learning to diagnose covid-19 in lung x-ray images with majority vote technique. 2020.

[Guo 2020] T. Guo. Heterogeneity aware federated learning. *A Major Qualifying Project Worcester Polytechnic Institute*, 2020.

[Haque *et al.* 2021] AKM Haque, Tahmid Hasan Pranto, Abdulla All Noman, and Atik Mahmood. Insight about detection, prediction and weather impact of coronavirus (covid-19) using neural network. *arXiv preprint arXiv:2104.02173*, 2021.

[He *et al.* 2020] Chaoyang He, Murali Annavaram, and Salman Avestimehr. Towards non-iid and invisible data with fednas: federated deep learning via neural architecture search. *arXiv preprint arXiv:2004.08546*, 2020.

[Heidari *et al.* 2020] Morteza Heidari, Seyedehnafiseh Mirniaharikandehei, Abolfazl Zargari Khuzani, Gopichandh Danala, Yuchen Qiu, and Bin Zheng. Improving the performance of cnn to predict the likelihood of covid-19 using chest x-ray images with preprocessing algorithms. *International journal of medical informatics*, 144:104284, 2020.

[Kandel and Castelli 2020] Ibrahem Kandel and Mauro Castelli. The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset. *ICT express*, 6(4):312–315, 2020.

[Kandel *et al.* 2020] Ibrahem Kandel, Mauro Castelli, and Aleš Popovič. Comparative study of first order optimizers for image classification using convolutional neural networks on histopathology images. *Journal of imaging*, 6(9):92, 2020.

[Kassania *et al.* 2021] Sara Hosseinzadeh Kassania, Peyman Hosseinzadeh Kassanib, Michal J Wesolowskic, Kevin A Schneidera, and Ralph Detersa. Automatic detection of coronavirus disease (covid-19) in x-ray and ct images: a machine learning based approach. *Biocybernetics and Biomedical Engineering*, 41(3):867–879, 2021.

[Khan *et al.* 2022] Azrin Khan, Rachael Garner, Marianna La Rocca, Sana Salehi, and Dominique Duncan. A novel threshold-based segmentation method for quantification of covid-19 lung abnormalities. *Signal, Image and Video Processing*, pages 1–8, 2022.

[Kumar *et al.* 2022] Vijay Kumar, Anis Zarrad, Rahul Gupta, and Omar Cheikhrouhou. Cov-dls: Prediction of covid-19 from x-rays using enhanced deep transfer learning techniques. *Journal of Healthcare Engineering*, 2022, 2022.

[Liu *et al.* 2020] Boyi Liu, Bingjie Yan, Yize Zhou, Yifan Yang, and Yixian Zhang. Experiments of federated learning for covid-19 chest x-ray images. *arXiv preprint arXiv:2007.05592*, 2020.

[MacMillan and Bensinger 2019] D MacMillan and G Bensinger. Google almost made 100,000 chest x-rays public—until it realized personal data could be exposed. *Washington Post*, 2019.

[Masoudi *et al.* 2021] Samira Masoudi, Stephanie AA Harmon, Sherif Mehralivand, Stephanie M Walker, Harish Raviprakash, Ulas Bagci, Peter L Choyke, and Baris Turkbey. Quick guide on radiology image pre-processing for deep learning applications in prostate cancer research. *Journal of Medical Imaging*, 8(1):010901, 2021.

[Masud 2022] Mehedi Masud. A light-weight convolutional neural network architecture for classification of covid-19 chest x-ray images. *Multimedia systems*, pages 1–10, 2022.

[McMahan *et al.* 2017] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. pages 1273–1282, 2017.

[Mukherjee *et al.* 2021] Himadri Mukherjee, Subhankar Ghosh, Ankita Dhar, Sk Md Obaidullah, KC Santosh, and Kaushik Roy. Deep neural network to detect covid-19: one architecture for both ct scans and chest x-rays. *Applied Intelligence*, 51(5):2777–2789, 2021.

[Nicola Flor and Brambilla 2021] Anna Paola Savoldi-Renato Vitale Giovanni Casazza Paolo Villa Nicola Flor, Lorenzo Saggiante and Anna Maria Brambilla. Diagnostic performance of chest radiography in high covid-19 prevalence setting: experience from a european reference hospital. *Emerg Radiol. Jul 28(5)*, page 877–885, 2021.

[Rieke *et al.* 2020] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman, Klaus Maier-Hein, et al. The future of digital health with federated learning. *NPJ digital medicine*, 3(1):1–7, 2020.

[Salman *et al.* 2020] Fatima M Salman, Samy S Abu-Naser, Eman Alajrami, Bassem S Abu-Nasser, and Belal AM Alashqar. Covid-19 detection using artificial intelligence. 2020.

[Tan *et al.* 2021] Wenjun Tan, Pan Liu, Xiaoshuo Li, Yao Liu, Qinghua Zhou, Chao Chen, Zhaoxuan Gong, Xiaoxia Yin, and Yanchun Zhang. Classification of covid-19 pneumonia from chest ct images based on reconstructed super-resolution images and vgg neural network. *Health Information Science and Systems*, 9(1):1–12, 2021.

[Tsang 2014] S Tsang. Review: Vggnet–1st runner-up (image classification). *Winner (Localization) in ILSVRC*, 2014.

[Tsiknakis *et al.* 2020] Nikos Tsiknakis, Eleftherios Trivizakis, Evangelia E Vassalou, Georgios Z Papadakis, Demetrios A Spandidos, Aristidis Tsatsakis, Jose Sánchez-García, Rafael López-González, Nikolaos Papanikolaou, Apostolos H Karantanas, et al. Interpretable artificial intelligence framework for covid-19 screening on chest x-rays. *Experimental and Therapeutic Medicine*, 20(2):727–735, 2020.

[UNCTAD 2020] UNCTAD. Data protection and privacy legislation worldwide. 2020.

[Yang *et al.* 2019a] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19, 2019.

[Yang *et al.* 2019b] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19, 2019.

[Yang *et al.* 2021a] Dandi Yang, Cristhian Martinez, Lara Visuña, Hardev Khandhar, Chintan Bhatt, and Jesus Carretero. Detection and analysis of covid-19 in medical images using deep learning techniques. *Scientific Reports*, 11(1):1–13, 2021.

[Yang *et al.* 2021b] Dong Yang, Ziyue Xu, Wenqi Li, Andriy Myronenko, Holger R Roth, Stephanie Harmon, Sheng Xu, Baris Turkbey, Evrim Turkbey, Xiaosong Wang, et al. Federated semi-supervised learning for covid region segmentation in chest ct using multi-national data from china, italy, japan. *Medical image analysis*, 70:101992, 2021.

[Zawad *et al.* 2021] Syed Zawad, Jun Yi, Minjia Zhang, Cheng Li, Feng Yan, and Yuxiong He. Demystifying hyperparameter optimization in federated learning. 2021.

[Zhang *et al.* 2021] Weishan Zhang, Tao Zhou, Qinghua Lu, Xiao Wang, Chunsheng Zhu, Haoyun Sun, Zhipeng Wang, Sin Kit Lo, and Fei-Yue Wang. Dynamic-fusion-based federated learning for covid-19 detection. *IEEE Internet of Things Journal*, 8(21):15884–15891, 2021.