Integrating Position-Aware Neurons in Deep Learning Models for Effective Federated Medical Image Classification

Nursultan Makhanov
School of Engineering and Digital Sciences
Nazarbayev University

Submitted by Nursultan Makhanov, to the Nazarbayev University as a thesis for the degree of Doctor of Philosophy in Computer Science, May, 2025.

I certify that all material in this thesis as my own work has been identified and that any material that has previously been submitted and approved for the award of a degree by this or any other University has been acknowledged.

Signed: .....

Acknowledgements

The completion of this doctoral dissertation represents not only years of intense research and study but also the generous support and guidance of many individuals to whom I am deeply indebted.

I wish to express my profound gratitude to my supervisory committee. I thank Dr. Anh Tu Nguyen, my principal supervisor, for providing exceptional guidance, intellectual stimulation, and unwavering support throughout my doctoral journey. Your rigorous approach to research and commitment to excellence have been instrumental in shaping this thesis. I also thank Dr. Kok-Seng Wong, whose innovative perspectives challenged my thinking and expanded the horizons of my research. Your insightful feedback consistently elevated the quality of my work. I also thank Dr. Muhammed Fatih Demirci, whose methodological expertise and attention to detail significantly strengthened the analytical framework of this dissertation. I am privileged to have benefited from your collective wisdom and mentorship.

On a personal note, I owe an immeasurable debt of gratitude to my wife Alina Makhanova. Your unconditional love, patience, and encouragement sustained me through the most challenging moments of this endeavor. You celebrated my successes and provided comfort during setbacks, while making countless sacrifices to support my academic pursuits. This achievement is as much yours as mine.

I also acknowledge my appreciation to the School of Engineering and Digital Sciences, Computer Science department staff, for their administrative assistance, and to the Nazarbayev University for providing the resources necessary for conducting this research.

Finally, I am grateful to the funding bodies and institutions that financially supported this research, making this academic journey possible.

Abstract

Privacy regulations create significant barriers to accessing distributed medical imaging datasets, impeding collaborative efforts to improve diagnostic accuracy for respiratory diseases. This thesis presents a novel Federated Learning (FL) framework that enables collaboration between healthcare institutions without compromising data privacy. The research makes three key contributions. First, it develops a comprehensive benchmarking framework to evaluate FL performance in medical image classification. This framework systematically compares various deep learning (DL) architectures, including Convolutional Neural Networks (CNNs), Transformers, and hybrid models, under challenging real-world conditions such as non-IID (non-independent and identically distributed) and imbalanced datasets. Second, the research introduces CoAtPENet, a hybrid model that enhances the CoAtNet architecture by incorporating Position-Aware Neurons (PANs). This integration specifically addresses the neuronal alignment issues that commonly occur during model aggregation in FL, improving model consistency across distributed training. Third, extensive empirical validation using three publicly available chest X-ray datasets demonstrates that the proposed approach achieves classification performance comparable to centralized training methods in both multi-class and multi-label classification scenarios. The results confirm that CoAtPENet effectively handles data heterogeneity while maintaining diagnostic accuracy. By enabling secure collaboration between healthcare facilities with protected datasets, this research advances both the theoretical understanding and practical application of FL in medical imaging, ultimately supporting a more accurate diagnosis of respiratory diseases without compromising patient privacy.

Contents

List of Figures

List of Tables

1. Introduction

1.1 Background and Evolution of AI in Medical Imaging

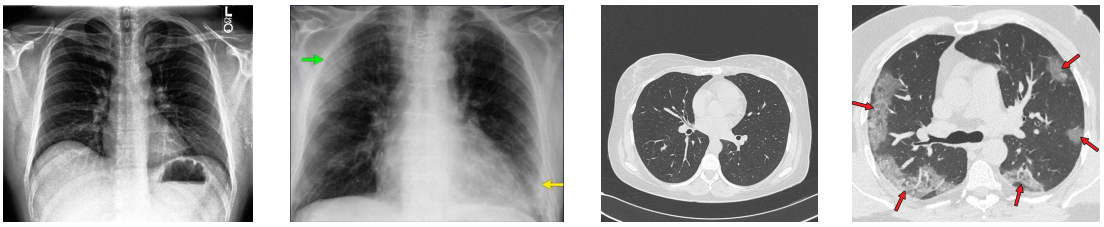
The use of artificial intelligence (AI) within the realm of medical imaging has undergone significant evolution over the last few decades. Early AI systems were primarily rule-based, relying on predefined criteria to detect abnormalities in medical images. However, these systems lacked the adaptability needed for complex diagnostic tasks. The shift to machine learning (ML) introduced data-driven models that could learn from examples, marking a major milestone in the field. With the development of Convolutional Neural Networks (CNNs) in the 1990s, AI began to achieve state-of-the-art performance in image classification tasks. More recently, DL has further advanced the field, with models such as CNNs and Vision Transformers being applied to a wide range of diagnostic tasks, from tumor detection to disease classification. DL models have become the backbone of Medical Image Analysis (MIA). These models work by processing raw image data through multiple layers of artificial neurons, allowing them to learn complex patterns and features. In the context of medical imaging (Fig. 1.1), DL models can identify intricate structures such as tumors or abnormalities in organs, which are often difficult to discern by traditional methods. Unlike traditional image processing techniques, which rely on manually designed features, DL models automatically extract features from the data, offering higher accuracy and generalization capabilities. Moreover, DL models can learn from large and diverse datasets, making them more adaptable to different medical contexts, leading to more robust and scalable diagnostic solutions.

MIA has become a vital component of clinical diagnosis and patient management, playing a crucial role in modern healthcare. Using sophisticated computational tools, MIA plays an essential role in improving the detection and treatment of numerous diseases, including respiratory diseases such as pneumonia, COVID-19, and chronic obstructive pulmonary disease (COPD). For instance, the European Respiratory Society [24] highlights that respiratory conditions, such as tuberculosis, asthma, and

Figure 1.1: Examples of AI helping to detect brain tumors from MRI scan.

cancer, rank as the third leading cause of death globally. Since its emergence in 2019, COVID-19 has significantly impacted many countries and is classified as a viral lung infection, sharing similarities with traditional pneumonia. By offering data-driven, objective insights, MIA supports radiologists and clinicians in creating customized treatment strategies and enhancing surgical precision. Early detection of these conditions is critical, as it increases the chances of timely interventions, ultimately improving patient outcomes. In particular, with the global impact of the COVID-19 pandemic, the need for scalable, accurate, and privacy-conscious diagnostic tools has become more apparent. MIA, powered by deep learning (DL) algorithms, offers automated solutions to complex diagnostic tasks by analyzing vast amounts of X-ray (Fig. 1.2(a,b)) and CT (Fig. 1.2(c,d)) scan data efficiently, which would otherwise require considerable time and effort from skilled radiologists.

DL models offer significant advantages in automating these complex tasks as they have shown tremendous results in image classification [39, 35, 69, 16], speech recognition [33, 82], object detection [91, 37, 10], and many other tasks in the last decade. The integration of DL into MIA has revolutionized the field, offering not only improvements in diagnostic accuracy but also personalized patient management. DL models, particularly convolutional neural networks (CNNs), are adept at recognizing patterns in medical images, which aids in early disease detection and enhances decision-making processes. Researchers are actively focusing on the use of computer



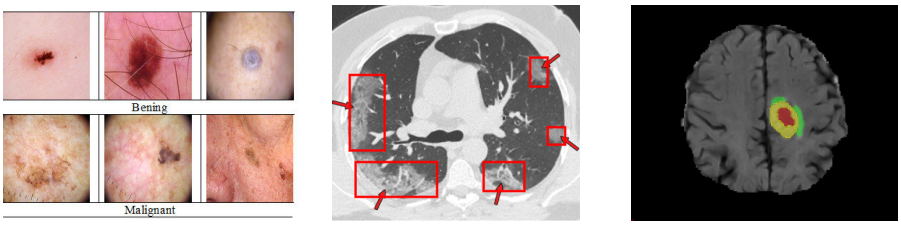
(a) Healthy X-ray (b) COVID-19 X-ray (c) Healthy CT (d) COVID-19 CT

Figure 1.2: Examples of healthy and COVID-19 infected X-rays and CT scans.

vision (CV) techniques for tasks such as disease classification [113, 79] (Fig. 1.3(a)), lung object detection [123] (Fig. 1.3(b)), and disease segmentation [34] (Fig. 1.3(c)). Using deep neural networks, these studies have shown notable success in diagnosing chest-related conditions. DL models can help detect lung abnormalities or cancerous lesions, facilitating more precise surgical planning and targeted treatment strategies. These models also provide clinicians with the ability to tailor treatment plans based on the patient's unique disease progression, improving overall healthcare outcomes and reducing recovery times.

1.2 Challenges in Medical Image Analysis

Despite advances, MIA faces several challenges that affect the development and deployment of AI models. Data quality is a critical issue, as medical images can vary in resolution, contain artifacts, or be affected by noise, which impacts the accuracy of DL models. Furthermore, variability in imaging protocols across different institutions can introduce inconsistencies in training data, making it harder for models to generalize across diverse settings. Inter-observer variability, where different radiologists provide varying interpretations of the same image, further complicates the labeling process, introducing bias into the datasets used to train AI models. These challenges underscore the need for more standardized imaging protocols and



(a) MIA classificaiton (b) MIA detection (c) MIA segmentation

Figure 1.3: Examples of Skin cancer classificaiton, COVID-19 detection, and Brain tumor segmentation.

more robust AI models capable of handling data variability in real-world applications.

Centralized training of DL models faces several technical hurdles, particularly for organizations with limited resources. The significant storage demands make it unsuitable for such settings. In multi-device environments, communication overheads lead to delays and potential bottlenecks during data transmission to a central server. Moreover, as datasets grow larger or models become more complex, scalability challenges emerge, rendering centralized approaches inefficient for computationally intensive tasks. Privacy concerns further complicate centralized training, as transmitting sensitive data to a central server often violates regulatory standards. For example, compliance with HIPAA [76] and GDPR [104] restricts the sharing of medical images containing personal health information, exacerbating issues such as data scarcity.

As AI becomes more integrated into healthcare care, ethical and regulatory considerations take greater importance. In the context of medical image analysis, HIPAA and GDPR impose strict guidelines on how patient data is handled. Ensuring compliance with these regulations is essential for maintaining patient privacy and trust. However, concerns about AI bias and fairness remain pressing. If AI models are trained on biased datasets, they may inadvertently perpetuate healthcare disparities, leading to unequal outcomes between different patient populations. Addressing these concerns requires careful consideration of the data used to train AI models, as well as ongoing efforts to improve fairness and transparency in AI systems.

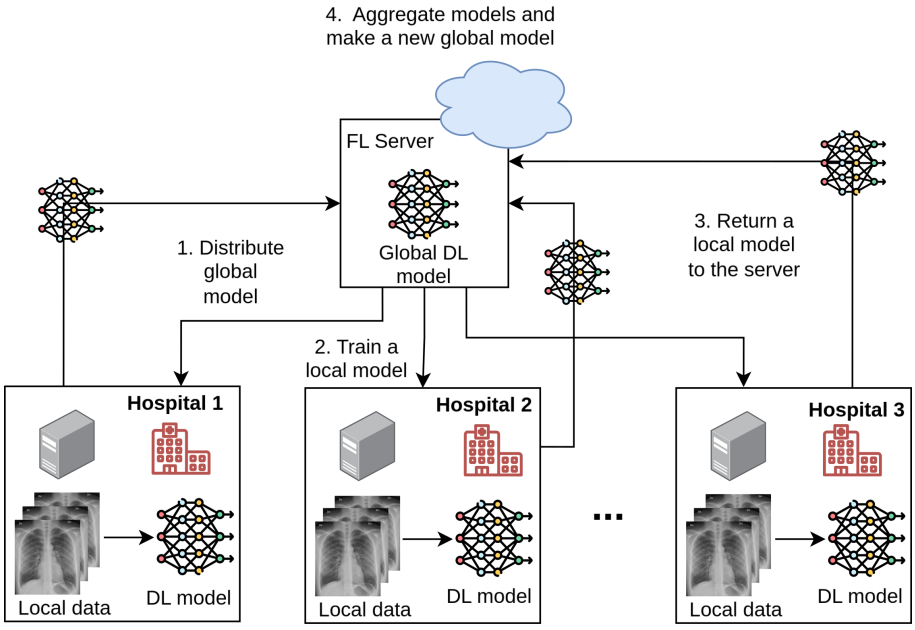


Figure 1.4: FL Examples applied to MIA.

1.3 Federated Learning for Medical Image Analysis

FL [68] addresses the challenges of data sharing while preserving privacy. This approach (Fig. 1.4) enables decentralized institutions to collaboratively train a DL model without transferring their private data, safeguarding patient confidentiality. Unlike centralized learning methods, FL reduces both storage requirements and communication burdens by localizing the training process. Rather than sending raw data to a central server, only model updates are exchanged, significantly cutting bandwidth usage and storage needs. Using the computational power of local devices, FL eliminates the dependency on a large centralized infrastructure, leading to more cost-effective operations.

The asynchronous structure of FL further enhances its efficiency, allowing devices to participate flexibly and reducing communication overhead during training. Numerous studies have proposed strategies for optimizing FL in various contexts [50, 58, 55]. For example, the FedAvg algorithm [50] combines local models on a central server through coordinate-based averaging to create a robust global model. However, FedAvg faces challenges when dealing with highly imbalanced data and fails to account for the permutation-invariant characteristics of models, resulting in neuron misalignment. This misalignment can hinder the learning process, leading to slower training convergence, less efficient learning, and reduced generalizability of the model. Proper alignment of the neuron is essential to maximize the performance of the DL models and ensure effective learning.

The FedProx algorithm [55] was one of the first methods designed to address the challenges posed by non-IID data distributions in Federated Learning. Building on the FedAvg algorithm, FedProx incorporates a proximal term that constrains the influence of local variables on the global model. Although FedProx follows a similar model-averaging strategy as FedAvg, its aim is to enhance stability during training in heterogeneous settings. However, despite these improvements, FedProx does not resolve the issue of neuron misalignment. This issue arises when models aggregated from participating clients exhibit inconsistencies in neuron alignment after being averaged on a central server. This misalignment can negatively impact the learning process, leading to reduced model performance and effectiveness.

In recent years, the issue of neuron misalignment during FL averaging has gained significant attention from researchers. Techniques for neuron matching, such as those proposed in [126, 127, 111, 124], represent some of the initial efforts to address this challenge. Although these methods aim to solve the misalignment problem, they

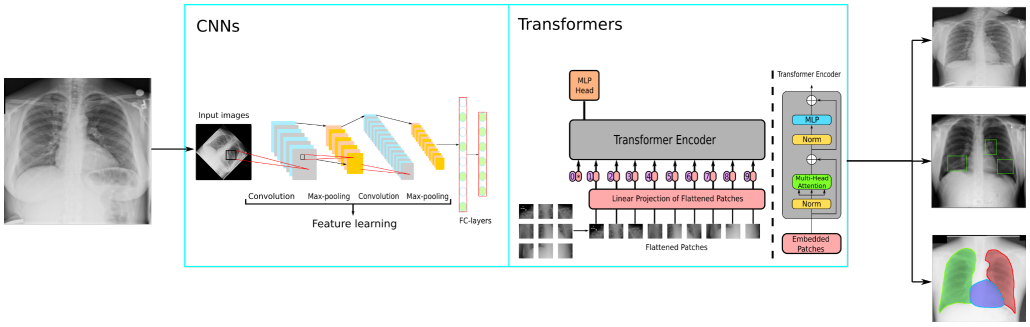


Figure 1.5: Examples of CNNs and Transformers applied to MIA.

often come with complexities in implementation or require extensive customization of neural networks. An alternative approach to tackle neuron misalignment is Federated Learning with Position-Aware Neurons (FedPANs) [59]. This method takes advantage of the permutation invariance property of DL models by introducing minimal modifications to the network architecture. A key advantage of FedPANs is its compatibility with a wide range of FL algorithms. However, a limitation of this approach is that it has primarily been applied to Convolutional Neural Network (CNN)-based architectures, leaving its potential for other model types relatively unexplored.

1.4 Deep Learning Models for Medical Image Analysis

Although CNNs (Fig. 1.5) excel in various tasks, their inability to effectively capture long-range dependencies can limit their performance in applications that require a broader understanding of the global context. To address this limitation, Vision Transformers (ViTs) [18] were introduced (Fig. 1.5), using self-attention mechanisms to model such dependencies. However, ViTs are computationally intensive, require significant memory, and lack the inductive biases inherent to CNNs. As a result, they require large datasets and substantial computational resources to achieve optimal performance.

To address these limitations, hybrid models such as CoAtNet [16] have been developed, merging the spatial encoding strengths of CNNs with the global context modeling capabilities of ViTs. CoAtNet has been shown to be highly effective in capturing local and global features, making it particularly well suited for MIA [46, 74]. This innovative approach provides an efficient and accurate solution for real-world medical applications that balances performance with resource constraints.

In recent years, the integration of deep learning models with FL has gained significant attention to automating lung disease detection and analysis [83, 52, 66]. However, comparing FL approaches and assessing their performance in medical image analysis (MIA) is challenging due to inconsistent experimental settings across studies. Moreover, many existing methods struggle with the challenges posed by highly imbalanced and non-IID data distributions, especially in multi-label datasets. Implementing non-IID setups for such datasets remains a complex task.

Recent studies addressing multi-label datasets such as CheXpert [41] in the FL context include [11, 83]. However, [11] utilized the entire dataset with only a small number of participants, which does not align well with real-world FL scenarios requiring a larger pool of clients and selective participant engagement. However, the work in [83] did not specify the FL method employed, limiting its reproducibility. Furthermore, neither of these studies examined the interaction between different FL techniques and various DL architectures. Yang et al. [121] conducted a detailed analysis of FL approaches on medical datasets but focused exclusively on a single CNN-based model, restricting the scope of fair and comprehensive evaluations.

1.5 Motivation, Research Questions and Objectives

Motivation of this work is to address the critical challenges in federated learning for medical image analysis, particularly focusing on neuron misalignment, data heterogeneity, and the need for efficient deep learning architectures that can perform well in diverse healthcare institutions while preserving patient privacy. By developing more robust federated learning approaches and hybrid models specifically designed for medical imaging tasks, this research aims to address several critical gaps in the current understanding of federated learning for medical image analysis. Based on the challenges and limitations identified in the literature, we formulate the following research questions:

1. How do different deep learning architectures (CNNs, Transformers, and hybrid models) perform in federated learning settings for medical image classification?

2. Can the integration of Position-Aware Neurons (PANs) into hybrid models like CoAtNet effectively address the neuron misalignment problem in federated learning?

3. How do data distribution characteristics (IID vs. non-IID, balanced vs. imbalanced) affect the performance of federated learning models in medical image analysis?

4. What is the impact of varying client participation and data partitioning strategies on the effectiveness of federated learning for medical image classification?

These research questions are motivated by several factors. First, the lack of standardized benchmarks for comparing different FL approaches in medical image analysis makes it difficult to determine which architectures are the most effective. Second, the neuron misalignment problem remains a significant challenge in FL, particularly for complex models and heterogeneous data distributions common in medical applications. Third, real-world medical data is often imbalanced and non-IID across institutions, yet the impact of these characteristics on FL performance is not well understood. Finally, practical FL deployments must consider varying levels of client participation and data availability, which requires an understanding of how these factors influence model performance.

To address the first questions, this study evaluates the performance of various CNN, Transformer, and hybrid models integrated with FL algorithms for medical image classification using three benchmark datasets. We compared traditional architectures like DenseNet121 and ResNet50 with newer attention-based models like MobileViT and hybrid models such as CoAtNet. As DL architectures continue to evolve rapidly, the need for a comprehensive benchmarking framework becomes increasingly evident. Our evaluation demonstrates meaningful comparisons between models, addressing the current limitation of standardized benchmarks for reliable conclusions about performance and efficiency in FL contexts for medical imaging.

Our research directly addresses the second question by demonstrating how Position-Aware Neurons (PANs) effectively mitigate the neuron misalignment problem in federated learning. We propose CoAtPENet, which enhances CoAtNet by embedding PANs throughout both the convolutional and the attention layers. As detailed in Chapter 4, PANs introduce position-specific information to neurons through additive and multiplicative mechanisms (Eqs. 4.1 and 4.2), creating a consistent neuron ordering that persists during model aggregation. Our experimental results in Chapter 6 quantitatively demonstrate the effectiveness of this approach, with CoAtPENet showing significantly reduced performance fluctuations compared to standard architectures. For example, in Figure 6.6, PAN equipped models exhibit smoother learning curves and up to 6-7% higher accuracy in non-IID settings. This

stability is particularly evident in later training rounds (40-100), where standard models often show significant accuracy oscillations while CoAtPENet maintains consistent performance. The PAN enhanced architecture effectively preserves learned feature representations during the averaging process, as evidenced by the improved convergence rates and higher final accuracy across all tested datasets. These results confirm that PANs successfully address the challenge of neuron misalignment by maintaining neuron consistency among federated clients, thus reducing the negative impact of weight averaging during aggregation.

To address the third question, we conducted experiments to evaluate the impact of data distribution characteristics (IID vs. non-IID, balanced vs. imbalanced) on the performance of FL models in medical image analysis. This includes simulating real-world conditions by constructing varied data partitions across clients that reflect typical variations in labels, features, and sample sizes in medical image analysis. Our results in Section 6.1 demonstrate that all models experience performance degradation in non-IID settings, with CNN-based architectures showing the most significant drops (up to 25% decrease in accuracy for DenseNet121), while hybrid models like CoAtPENet maintain more robust performance (only 7-8% decrease). For multi-label classification tasks on CheXpert and MIMIC-CXR datasets, we observed that imbalanced data distributions reduced mean AUROC scores by 3-5% across all architectures, with CoAtPENet showing the smallest performance gap between balanced and imbalanced settings.

To address the fourth research question, we conducted a comprehensive analysis of how varying client participation and data partitioning strategies affect the effectiveness of federated learning. In Sections 6.3 and 6.4 of Chapter 6, we systematically investigated the impact of changing both the number of clients (from 2 to 100) and the percentage of participants per round (from 10% to 100%). Our findings reveal that increasing the number of clients generally leads to a decrease in performance due to the greater fragmentation and heterogeneity of the data, the accuracy falling approximately 15% when scaling from 2 to 20 clients in the CovidX dataset (Figure 6.18). For multi-label datasets, we observed similar trends in Figures 6.19 and 6.20, with performance declining as client numbers increased. Regarding participation rates, our experiments show that while full client participation typically yields the best results, CoAtPENet maintains robust performance even with reduced participation, demonstrating only a 2-3% drop in AUROC when participation decreases from 100% to 10% (Figures 6.22 and 6.23). We also compared different data partitioning approaches, implementing Latent Dirichlet Allocation (LDA) for multi-class data and a custom splitting function for multi-label data. Our custom partitioning

strategy for multi-label data proved particularly effective, ensuring sufficient training samples per client while realistically simulating the imbalanced nature of medical data distributions. These findings provide valuable information for the deployment of FL systems in healthcare settings, where the number of participating institutions and data distribution characteristics significantly impact model performance.

1.6 Contributions and Dissertation Structure

We outline our contributions as follows.

- We introduce a comprehensive FL framework designed for medical image classification, which typically involves sensitive patient data distributed across multiple healthcare institutions. This framework facilitates the training of collaborative models while safeguarding data privacy and security.
- We examine the effects of several DL models, such as DenseNet121 [39], ResNet50 [35], and attention-based MobileViT [69]. Specifically, we investigate the performance of the hybrid CoAtNet model [16] in medical image classification tasks.
- We implement a straightforward approach where clients are assigned a fixed number of images in each round, with constraints, to tackle the challenge of managing multi-label datasets.
- We present the CoAtPENet model, which merges CoAtNet with Position-Aware Neurons (PANs) in the FL framework. We also conduct an empirical study to evaluate CoAtPENet with various FL algorithms.
- We carry out thorough experiments on benchmark datasets under realistic FL conditions, including various client setups, different participant percentages, IID and non-IID data distributions, balanced and imbalanced data configurations, and PANs both activated (ON) and deactivated (OFF).

The dissertation is structured as follows.

- Chapter 2 reviews the literature on medical image classification and FL.
- Chapter 3 revisits the popular FL algorithms and two classification problems for medical imaging.
- Chapter 4 outlines the methodology, data partitioning strategies, and architecture of the proposed CoAtPENet model.

- Chapter 5 shows the experimental setup and describes the datasets used to train and test multiple DL models in the FL setups.
- Chapter 6 presents the results of the experiments conducted, comparing CoAtPENet with other models under different FL setups.
- Chapter 7 discusses the limitations of the study and potential areas for future research.
- Chapter 8 concludes the dissertation with a summary and suggestions for future research.

2. Related Works

2.1 Medical Image Analysis.

The general workflow for MIA is illustrated in Fig. 2.1. Initially, raw DICOM files undergo several preprocessing steps, including resizing, rotating, scaling, and others. Once processed, these files are fed into DL models such as Transformers, CNNs, GANs, RNNs, which play a crucial role in feature extraction. By reducing dimensionality and automatically learning critical representations, these models facilitate tasks such as object classification, detection, and segmentation.

Image classification is a computer vision (CV) task that involves identifying objects within images or videos based on their distinct characteristics. Before the advent of DL algorithms, image classification was a significant challenge. However, since 2013, DL techniques have achieved impressive results in a variety of classification tasks. Image classification can now be applied in any domain that involves images, videos, or cameras. In particular, its use has expanded rapidly in the medical field [114, 81, 42, 79, 117].

Object detection entails the identification and localization of objects in images or videos by enclosing them within bounding boxes. This task is vital for various applications, including autonomous vehicles and video surveillance. Object detection

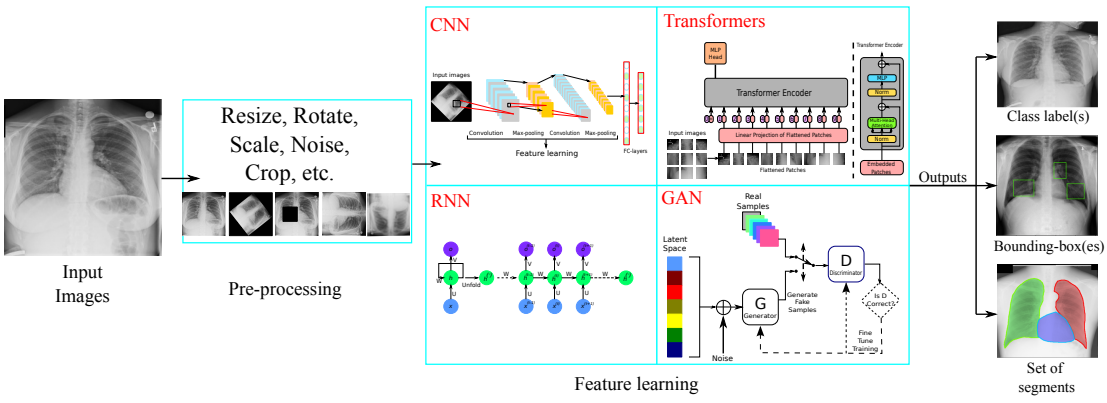


Figure 2.1: General Workflow of Medical Image Analysis

algorithms are typically categorized into one-stage and two-stage methods. One-stage approaches, such as YOLO [90], prioritize speed and simplicity, making them suitable for real-time applications. In contrast, two-stage methods, including Fast R-CNN [28] and Faster R-CNN [92], achieve higher accuracy, but often come with increased computational requirements and slower processing times.

Image segmentation is the process of partitioning an image into meaningful segments by grouping pixels associated with the same object type. There are three primary categories: semantic segmentation [75], instance segmentation [64], and panoptic segmentation [49]. In semantic segmentation, multiple objects of the same class are considered collectively as one entity. In contrast, instance segmentation distinguishes among individual objects of the same class. For example, in chest radiography (CXR) analysis, semantic segmentation would involve separating the lungs and bones as unique segments. Panoptic segmentation merges both concepts, handling segmentation at the instance and semantic levels simultaneously.

Recent advances in CV have resulted largely from the advent of CNNs. These networks typically consist of an input layer that processes raw image pixels, a series of hidden layers where feature extraction and representation learning occur, and an output layer responsible for generating class predictions. Transfer learning involves reusing pre-trained DL models, originally developed for one domain, to efficiently address similar or related CV tasks, such as those encountered in the medical field. Although CNNs offer high accuracy with minimal human involvement, their training requires substantial data, significant time investment, and extensive computational resources.

Using multiple snapshots of the COVID-Net model, Tang et al. [103] applied an ensemble DL approach and reported a precision of 95%. Wang et al. [114] had introduced COVID-Net incorporating a projection-expansion-projection framework which achieved 93.3% accuracy in predicting COVID-19. Meanwhile, Pham et al. [81] examined 16 pre-trained CNNs to detect COVID-19 on CT scans. Chowdhury et al. [14] focused on distinguishing COVID-19 pneumonia from normal and viral pneumonia on chest radiographs, achieving 99. 7% accuracy. Furthermore, Ahuja et al. [4] employed wavelet transform augmentation combined with a pre-trained ResNet18 model, achieving 99. 4% accuracy in the COVID-19 classification based on CT.

Recurrent Neural Networks (RNNs) constitute a class of artificial neural networks (ANNs) adept at handling sequential data by preserving internal states that capture contextual cues from previously encountered input. Despite this strength, standard RNNs encounter difficulties when processing very long input sequences, often resulting

in the problem of vanishing gradients. To mitigate this, the Long Short-Term Memory (LSTM) algorithm [47] was introduced as a refined variant of RNN capable of retaining information for extended periods. Although LSTMs enhance memory retention, they can still struggle with extremely long sequences.

In efforts to apply RNN/LSTM architectures to detect COVID-19, various investigations were examined. For example, Islam et al. [42] developed a combined CNN-LSTM method to predict COVID-19, where CNNs extracted features and LSTMs performed classification. Similarly, Sedik et al. [96] adopted a CNN-LSTM framework to identify COVID-19 in CT and X-ray images. Extending this idea, Pustokhin et al. [84] proposed the RCAL-BiLSTM model, incorporating a ResNet-based Class Attention Layer with a Bidirectional LSTM for COVID-19 detection. Their model achieved 94.88% accuracy, underscoring the potential of hybrid architectures to improve diagnostic performance. For nearly a decade, CNNs have reigned as the primary architecture in computer vision. Although alternatives were explored, CNNs continued to dominate as their depth and complexity expanded. A significant change occurred when Dosovitskiy et al. [18] introduced the Vision Transformer (ViT), a Transformer-based model that demonstrated the feasibility of image classification without relying on CNNs. The strength of ViT lies in its ability to handle long input sequences; however, this advantage comes with the drawback that ViT models generally need larger datasets than CNNs to achieve optimal performance.

In the context of COVID-19 pneumonia detection, Park et al. [79] applied ViT after pre-training the backbone network with PCAM, and then fine-tuning the entire architecture. Meanwhile, Sriram et al. [100] introduced a Transformer-based model enhanced through self-supervised pre-training with Momentum Contrast Learning, achieving AUC values of 0.786 and 0.848 in the prediction of adverse events and mortality, respectively.

Generative Adversarial Networks (GANs) [32] are generative deep learning models frequently employed in unsupervised settings. A GAN comprises a generator, responsible for creating synthetic samples, and a discriminator, which determines whether these samples appear authentic. Although GANs excel at producing artificial data when the available dataset is limited, radiological validation of the generated samples remains essential in medical imaging.

In the realm of COVID-19 analysis, Yadav et al. [117] introduced an unsupervised GAN-based approach combined with a support vector classifier (SVC) to examine chest X-ray images. Here, the GAN acted as a feature extractor, while its output features were then classified using SVC and logistic regression. Other investigators [61, 30, 27] applied GAN-driven data augmentation to produce additional CT images

for COVID-19 detection, achieving accuracy scores of 93%, 99.22%, and 99.60%, respectively. Furthermore, Quan et al. [85] developed XPGAN to improve the classification of COVID-19 in X-ray images by augmenting CT scans, resulting in an F1 score of 0.823. Currently, research focusing on localizing COVID-19 pneumonia within X-rays and CT scans remains limited, encouraging further discussion of object detection methods and their potential applications in the location of COVID-19-associated pneumonia regions.

R-CNN [29], a basic two-stage object detection method, addresses both object detection and segmentation by generating candidate regions through Selective Search [108], classifying these proposed regions, and predicting labels as well as bounding boxes. Despite its accuracy, R-CNN's slow processing speed motivated the creation of Fast R-CNN [28], which employs a convolution-based sliding window to classify all proposed regions more efficiently. Building on this, Ren et al. introduced Faster R-CNN [92], leveraging a convolutional network to generate multiple object proposals directly. However, effectively handling overlapping bounding boxes remains a persistent challenge.

These methodologies have been applied to medical imaging tasks. Yao et al. [123] utilized Faster R-CNN to detect pneumonia in CXR images, achieving mean average precision (mAP) scores of 39.23% in the RSNA dataset and 38.02% on ChestX-ray14. Li et al. [56] incorporated Faster R-CNN with reverse learning (through a gradient reversal layer) in their NIA-Network to identify COVID-19 infections on CT scans, reaching an accuracy of 92.1%. Furthermore, Yang et al. [122] proposed a multi-deep learner approach, combining Mask R-CNN with ResNet50, Mask R-CNN with ResNet101, and Faster R-CNN with ResNet101 for the detection of pneumonia in CXR images, obtaining a precision rate of 96% for bounding box predictions.

One-stage object detection algorithms like YOLO [90] focus on real-time processing. YOLO partitions the input image into a grid structure, where each cell predicts one or more bounding boxes with associated confidence scores. After these bounding boxes are established, class probabilities are computed to determine which objects they contain. As an example, Al-antari et al. [6] employed YOLO to detect COVID-19 in X-ray images.

For image segmentation tasks, Mask R-CNN [36] stands as a state-of-the-art solution. In addition to detecting objects, it also produces high-quality segmentation masks for each identified instance. Among its key advantages are ease of training, superior accuracy compared to other segmentation algorithms, efficiency due to its foundation on Faster R-CNN, and adaptability to a wide range of tasks.

In practical applications, Nayyar et al. [72] implemented a Mask R-CNN-based

object detector to identify X-ray images containing pneumonia or other lung diseases, achieving an Intersection over Union (IoU) score of 0.155. Similarly, Ramesh et al. [87] used Mask R-CNN to segment COVID-19 lung lesions in chest radiographs, reaching a maximum IoU of 0.81. Another study, conducted by Wu et al. [116], utilized an encoder-decoder architecture for segmentation, reporting a Dice score of 0.783 and an IoU score of 0.665.

U-Net [94] is a deep learning model specifically designed to segment infected areas within biomedical images. Ronneberger et al. highlighted that its encoder-decoder structure, rooted in CNN architectures, enables effective feature extraction and accurate segmentation outcomes. In addition, the U-Net architecture balances global and local contextual information, although the internal bottleneck can slow the learning process.

Based on U-Net's principles, Hasan et al. [34] adapted a DenseNet-based U-Net framework to delineate COVID-19-infected regions in X-ray images, obtaining an average IoU of 0.90 and a Dice coefficient of 0.92. In another example, Munusamy et al. [71] introduced FractalCovNet to segment infected areas on CT scans, reporting a mean Absolute Error (MAE) of 0.064. Zhou et al. [132] integrated attention mechanisms into a U-Net-based COVID-19 segmentation algorithm, achieving a Dice score of 83.1% and a Hausdorff distance of 18.8 on 473 CT scans. Furthermore, Zhang et al. [129] combined Dense GAN with a U-Net enhanced by a multilayer attention mechanism (MLA) to isolate lung lesions in COVID-19 CT scans, with U-Net handling segmentation while Dense GAN facilitated data augmentation.

Medical image classification is an essential component of the broader image classification landscape, involving the application of machine learning and deep learning techniques to extract, learn, and categorize features in medical images. By leveraging these methods on modalities like CT or X-ray scans, it becomes possible to produce results more rapidly and at a lower cost than traditional diagnostic techniques. This efficiency has led to a surge of interest in identifying chest-related conditions—such as COVID-19, pneumonia, emphysema, COPD, and chest cancer—through automated DL-based strategies. In particular, COVID-19, as a recent and highly impactful outbreak, has prompted substantial research efforts to create advanced models that streamline its detection and diagnosis.

Historically, DL models have relied on centralized data collection, gathering large datasets on a single server before training. In early research addressing conditions like COVID-19 and pneumonia, many approaches employed deep CNNs [113, 80, 14]. Yadav et al. [118], for instance, assessed various ML and DL models for pneumonia classification using chest radiographs. Some studies integrated CNNs with LSTM

networks to enhance COVID-19 detection capabilities [42, 96]. Others explored contrastive learning methods for improved feature discrimination, as seen with Azizi et al. [7], who utilized multi-instance contrast learning for classifying medical chest X-ray images. Additionally, Abbas et al. [2] demonstrated the utility of a pre-trained DeTraC CNN model in accurately detecting COVID-19 within X-ray images.

Recent advancements in the Vision Transformer (ViT) have opened new avenues for utilizing attention mechanisms, thus significantly enhancing image classification performance. Numerous studies have leveraged ViT for COVID-19 detection in medical imaging. For instance, Shome et al. [97] introduced a ViT-based architecture that integrates a custom MLP block, while Park et al. [79] combined ViT with a PCAM network to diagnose COVID-19.

Additionally, recent research [86, 106, 57] has examined hybrid models blending CNNs and Transformers for medical image analysis (MIA). Raj et al. [86], for example, developed StrokeViT, which merges CNNs and ViT to enhance both slice-level and patient-level predictions in brain stroke classification by incorporating local feature extraction and capturing long-range dependencies. Meanwhile, Thon et al. [106] employed the Convolutional Vision Transformer (ConViT) to detect COVID-19 in lung CT scans, analyzing how image resolution and the number of attention heads affect model performance. In another study [57], the proposed multimodal medical image fusion model leverages a CNN module for detailed texture extraction and a Transformer module to capture pixel intensity distributions.

2.2 General Federated Learning Algorithms

FL is a recent development that allows training in a distributed fashion for several edge devices. Classical FedAvg [68] is the first and most widely used FL algorithm that distributes the global model to participating clients and aggregates all models in the server. However, one of the important properties of an FL algorithm is how it can handle the non-IID data. For instance, FedProx [55] and FedBN [58] are the modified versions of FedAvg where one introduces a proximal term, and the other eliminates the Batch Normalization layer to tackle the model convergence problem in non-IID. FedOpt [89] is another FL algorithm that addresses the convergence problem by optimizing global aggregation parameters using optimizers such as Adam [48], Yogi [128] instead of simple parameter averaging.

2.3 Personalized Federated Learning Algorithms

Beyond the established FL algorithms, personalized Federated Learning (PFL) [102] offers solutions to address key challenges posed by heterogeneous data distributions. A notable PFL approach is Per-FedAvg [21], an adaptation of FedAvg that builds upon the Model-Agnostic Meta-Learning (MAML) framework [23]. Per-FedAvg leverages meta-learning for rapid adaptation of ML/DL models to limited datasets, enabling better personalization of client-specific models while retaining the aggregation principles of FedAvg. This approach was further enhanced in [101] by incorporating an l_2-norm loss, which balances local and global model performance.

Another essential PFL framework is FedMD [53], which integrates transfer learning and knowledge distillation to develop personalized models. Transfer learning within FL has also been extensively explored in the medical domain to enhance model personalization, as demonstrated in work such as FedHealth [13] and FedSteg [119]. Additionally, FedHeNN [67] presents an architecture-agnostic framework that enables clients with various model structures to participate in federated learning. FedHeNN achieves this by aligning instance-level representations through a proximal term, thereby improving the overall performance of ML/DL models.

2.4 Federated Learning with Permutation Invariance.

The element-wise averaging mechanism employed by state-of-the-art FL algorithms has limitations, particularly when handling non-iid data distributions [62, 12]. The alternative methodologies proposed in [126, 112] offer more robust solutions by taking advantage of the permutation invariance of neurons in ML/DL models. For example, Bayesian nonparametric learning [126] and Federated Matched Averaging (FedMA) [112] use layer-wise neuron permutation to align similar neurons across models before averaging. This approach enables a more structured combination of model parameters compared to traditional averaging methods.

Similarly, Singh et al. [99] introduced a layer-wise model fusion algorithm that utilizes optimal transport to align neurons and averages their parameters, demonstrating significant performance improvements over vanilla averaging. The Fed^2 algorithm [124] further addresses structural feature misalignment by employing feature-paired averaging techniques. Another method, FedPANs [59], improves FL performance by using Position-Aware Neurons (PANs) to integrate position-related values into neuron output, exploiting the permutation invariance property of ML/DL

models to improve the aggregation process.

2.5 Federated Learning in Medical Image Classification.

Several studies have examined the implementation of Federated Learning (FL) for medical chest image classification. For example, Park et al. [78] introduced the FESTA framework, which combines split learning and Vision Transformers (ViTs) to diagnose COVID-19 in X-ray images. Kumar et al. [52] proposed a blockchain-supported FL approach to detect COVID-19 on CT scans, where the blockchain ensures data integrity while FL enables collective model training. Furthermore, Feki et al. [22] used the FedAvg algorithm with the VGG16 and ResNet50 architectures to identify COVID-19 in X-ray images. Huang et al. [40] introduced GLoRIA, a framework that uses contrastive learning on attention-weighted image regions paired with text reports to derive global and local multimodal representations. Adnan et al. [3] demonstrated differentially private FL through DenseNet and MEM models, ensuring privacy-preserving histopathological image analysis. A comprehensive overview of FL's role in COVID-19 detection via X-ray images is provided in [73].

Integrating differential privacy (DP) algorithms, as described in [19], injects controlled noise into data or models to improve security. However, using DP with FL may slightly compromise testing accuracy compared to centralized approaches. Zhang et al. [131] introduced a dynamic fusion-based FL strategy for COVID-19 detection in X-rays and CT scans, maintaining data privacy and achieving 95% accuracy. Likewise, Kumar et al. [52] applied a blockchain-integrated FL framework with SegCaps and Capsule Networks, achieving 98.68% accuracy in a highly secure and privacy-focused setup.

Beyond the methods previously discussed, researchers have also investigated federated learning (FL) and transfer learning approaches for the COVID-19 classification. For example, Liu et al. [63] and Feki et al. [22] examined how the combination of FL with transfer learning techniques can enhance the detection of COVID-19 in chest X-ray images. In another line of work, Ulhaq et al. [109] introduced a theoretical framework that integrates differential privacy (DP) into FL to predict COVID-19 cases, ensuring robust and scalable results within a secure environment.

In terms of privacy-preserving strategies, Muftuouglu et al. [70] and Yuan et al. [125] applied DP to deep learning models to predict COVID-19 in CXR images, achieving promising results under privacy constraints. Furthermore, Eom et al.

[20] proposed a vector-based data anonymization model, outperforming point-based approaches in balancing data utility with privacy. Similarly, Iyer et al. [43] developed a spatial k-anonymity algorithm to anonymize geolocation data of COVID-19 patients, improving privacy protection without compromising the usefulness of the data for contact tracking efforts.

2.6 Self-supervised Learning and Few-shot Learning in Medical Image Analysis.

Medical data sharing is inherently restricted due to privacy concerns, which poses challenges related to insufficient training data. Researchers often lack access to large datasets needed to train effective predictive models. Unlike ImageNet, which contains a large collection of images suitable for training DL models, medical image datasets are comparatively small, partly due to the personal information embedded in DICOM files. To address this limitation, recent research has increasingly focused on self-supervised learning (SSL) and few-shot learning (FSL) techniques as potential solutions.

Self-supervised learning enables machines to learn from unlabeled data by extracting useful representations autonomously, reducing the dependency on human-annotated datasets. In medical imaging, the ultimate objective is to extract meaningful labels from the input data without manual intervention. A major advantage of SSL is its ability to significantly reduce the amount of labeled data required for model training. However, SSL models may encounter overfitting issues when working with extremely limited data.

Li et al. [60] addressed the difficulties posed by weak annotations and limited datasets in medical imaging through a weakly supervised framework enhanced by self-supervision and multiple instance augmentations. Meanwhile, Fung et al. [25] introduced self-supervised learning into the Inf-Net model to segment coronavirus lesions directly from raw CT images, resulting in SSInfNet, which surpassed both U-Net and Inf-Net in performance. Abbas et al. [1] also used self-supervised learning (SSL) to classify COVID-19 on chest radiographs, combining deep CNN, transfer learning, and clustering algorithms. Their approach achieved 99.8% accuracy on two extensive datasets utilizing unlabeled images. Lastly, Park et al. [77] combined SSL, Models Genesis, and the convolution block attention module (CBAM) to improve the diagnosis of COVID-19 in CXR, further illustrating the potential of SSL-based techniques in this domain.

Few-shot learning (FSL) is an emerging machine learning strategy aimed at training models from extremely limited datasets, thereby reducing the expenses tied to extensive data collection and labeling. Although FSL can deliver performance comparable to that of leading techniques, its sensitivity to subtle shifts in data distribution can limit broader applicability. Although research on FSL for COVID-19 prediction remains relatively sparse, several notable examples showcase its potential.

For example, Ma et al. [65] used FSL in conjunction with domain generalization and knowledge transfer to address segmentation tasks using small-scale COVID-19 datasets. In another study, Jadon and Shruti [44] adopted an FSL framework that uses Siamese networks and a contrastive loss function to identify COVID-19 in chest X-ray images, achieving an accuracy of 96.4%, a significant improvement over a logistic regression baseline (83%). Similarly, Shorfuzzaman et al. [98] combined a pre-trained VGG-16 encoder with Siamese networks and various n-shot learning methods for COVID-19 classification, reporting a precision of 95.6%, with sensitivity and specificity scores of 96% and 98%, respectively.

2.7 Current State of Chest Medical Image Classification.

The current state of medical image classification is undergoing a significant transformation, characterized by the increasing integration of DL methodologies to address critical diagnostic challenges across a spectrum of diseases, including infectious and neurological conditions. A prominent area of focus has been the application of advanced computational architectures to COVID-19 diagnostic imaging, driven by the urgent need for rapid and accurate detection during the pandemic. Studies by Tang et al. [103], Chowdhury et al. [14], Shome et al. [97], Pham et al. [80], and Thon et al. [106] have demonstrated the efficacy of DL strategies in analyzing chest X-ray and CT scan data, emphasizing the automation of image interpretation to overcome the limitations of traditional diagnostic methods. This trajectory extends beyond COVID-19, as exemplified by Raj et al. [86], who applied hybrid algorithms to CT scan analysis for stroke detection, underscoring the broader potential of these techniques in automating complex diagnostic processes. A consistent methodological theme in these studies is the deployment of sophisticated DL architectures, such as ensemble models and ViT, to improve diagnostic speed, accuracy, and automation.

However, the application of these powerful techniques is often hindered by data privacy concerns and the challenge of data scarcity, particularly in medical contexts.

FL has emerged as a promising solution to these issues, enabling collaborative model training across decentralized datasets without necessitating direct data sharing. This approach is particularly relevant in medical imaging, where patient privacy is paramount. Research by Liu et al. [63], Yang et al. [120], Feki et al. [22], and Park et al. [78] demonstrates the efficacy of FL in the diagnosis of COVID-19, with innovations such as the FLOP algorithm by Yang et al. [120] optimizing privacy through partial model parameter sharing and Park et al. [78] exploring Vision Transformers in split learning to reduce bandwidth. Furthermore, Huang et al. [40] address the broader challenge of limited labeled medical image datasets by leveraging radiology reports for multimodal representation learning, reducing the dependency on manual annotations.

Recent FL research has also focused on overcoming challenges related to data heterogeneity and non-iid distributions. Chakravarty et al. [11] proposed a hybrid CNN-GNN architecture to accommodate site-specific co-morbidity variations in chest radiograph analysis. Kulkarni et al. [51] introduced FedFBN to enhance FL performance on non-iid and partially labeled datasets by freezing batch normalization layers. Tayebi et al. [105] systematically evaluated the impact of training strategies and dataset characteristics on FL generalization capabilities, highlighting the benefits of collaborative training between diverse institutions. Gong et al. [31] contributed a privacy-preserving FL framework using unlabeled public data for one-way offline knowledge distillation. Collectively, these studies underscore the importance of customized FL methodologies to address data privacy constraints and heterogeneity, enhancing the generalizability and robustness of diagnostic AI models.

The trajectory of medical image classification reflects a paradigmatic shift toward more sophisticated computational epistemology, integrating advanced DL techniques to improve diagnostic precision and efficiency. FL plays a crucial role in this evolution, allowing collaborative model development while adhering to stringent data privacy requirements, facilitating the translation of cutting-edge research into clinical practice.

3. Background

3.1 Federated Learning

Federated Learning (FL) focuses on collaborative model training without centralizing data. Instead of aggregating all data on a single server, FL methods aim to minimize a weighted average of the local objective functions across multiple clients. Formally, given K clients, the FL objective is defined as:

$$\min_{\mathbf{w}} F(\mathbf{w}) = \sum_{i=1}^{K} \frac{n_i}{n} F_i(\mathbf{w}) \tag{3.1}$$

Here, n_i denotes the number of data samples held by the i-th client, n is the total number of samples across all clients, and F_i represents the local objective function for the i-th client. Equation 3.1 inherently addresses real-world conditions in which different clients may hold varying amounts of data, leading to non-IID distributions.

FedAvg, introduced by McMahan et al. [68], is the most widely adopted FL algorithm. In FedAvg, a central server distributes a global model to all participating clients. Each client trains the model locally on its private dataset for several epochs to preserve data privacy, and then returns the updated model weights to the server. The server aggregates these updates as:

$$\mathbf{W}_r^{global} = \frac{1}{k} \sum_{i=1}^{k} \mathbf{W}_{i,r}$$

This iterative process refines the global model over multiple rounds, gradually improving its robustness and generality without ever requiring centralized data storage.

FedProx [55] extends the FedAvg algorithm to better handle heterogeneity among clients. It modifies the local objective function to:

$$\min_{\mathbf{w}} h_k(\mathbf{w}, \mathbf{w}^t) = F_k(\mathbf{w}) + \frac{\mu}{2} \|\mathbf{w} - \mathbf{w}^t\|^2 \tag{3.2}$$

In this formulation, $F_k(\mathbf{w})$ is the local objective for the k-th client, $\mathbf{w}$ denotes the current model weights, $\mathbf{w}^t$ are the global model weights from the previous round, and μ is a proximal term. By incorporating the proximal term, FedProx restricts the magnitude of local updates, preventing overly aggressive shifts in model parameters. This modification helps stabilize the training process and, as the authors claim, improves convergence in the presence of non-IID data distributions.

3.1.1 Federated Learning Based on Permutation Invariance

FedPANs [59] is a federated learning approach that capitalizes on the permutation invariance property of CNN models. To enhance existing FL methods, the authors introduced Position-Aware Neurons (PANs), which integrate position-related values into neuron output. In Fig. 3.1, we illustrate the core concept of the FedPANs algorithm, which involves permuting neurons of a neural network in a FL setting. The top portion shows the permutation module turned "off," where the network's neurons are not altered. In contrast, the bottom part demonstrates that the permutation module is turned on, activating the permutation operation. By making neurons aware ("on") of their spatial positions, PANs help mitigate the risk of misalignment of neurons.

Two versions of PANs are defined as follows:

$$\mathbf{PAN}_{+} : h_l = f_l((\mathbf{W}_l \mathbf{h}_{l-1} + \mathbf{b}_l) + \mathbf{e}_l) \tag{3.3}$$

$$\mathbf{PAN}_{\odot} : h_l = f_l((\mathbf{W}_l \mathbf{h}_{l-1} + \mathbf{b}_l) \odot \mathbf{e}_l) \tag{3.4}$$

In Eq. 3.3 and Eq. 3.4, additive (+) and multiplicative ($\odot$) PANs are implemented by adding or multiplying the position encodings $\mathbf{e}_l$ in the neuron outputs, respectively. Here, h_l is the output of layer l, f_l denotes the activation function, and $l \in \{0, 1, \ldots, L\}$ is the layer index. The bias term is $\mathbf{b}_l$, while the convolution parameters $\mathbf{W}_l \in \mathbb{R}^{C_l \times w_l \times h_l \times C_{l-1}}$ are defined by the input/output channels (C_l, C_{l-1}) and the dimensions of the kernel (w_l, h_l).

The position encodings $\mathbf{e}_l$ are generated using a sinusoidal function similar to the method described in [110]:

$$\mathbf{PAN}_{+} : e_{l,j} = B \sin(2\pi T j / J) \in [-B, B] \tag{3.5}$$

$$\mathbf{PAN}_{\odot} : e_{l,j} = 1 + B \sin(2\pi T j / J) \in [1 - B, 1 + B] \tag{3.6}$$

In these equations, T and B represent the period and amplitude of the position encoding, respectively, and $j \in \{0, 1, \ldots, J-1\}$ denotes the position index.

Using PANs, the parameters can be aligned in a coordinate-based manner, facilitating a more effective parameter averaging between clients in a federated learning environment.

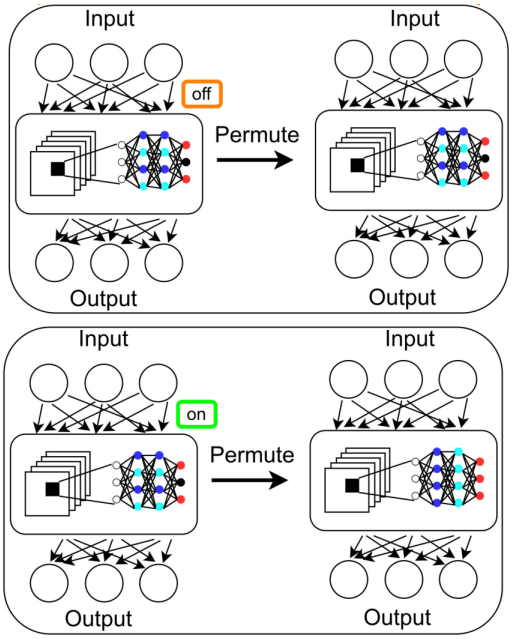


Figure 3.1: Illustration of FedPANs with PANs ON or OFF

3.2 Multi-class Classification

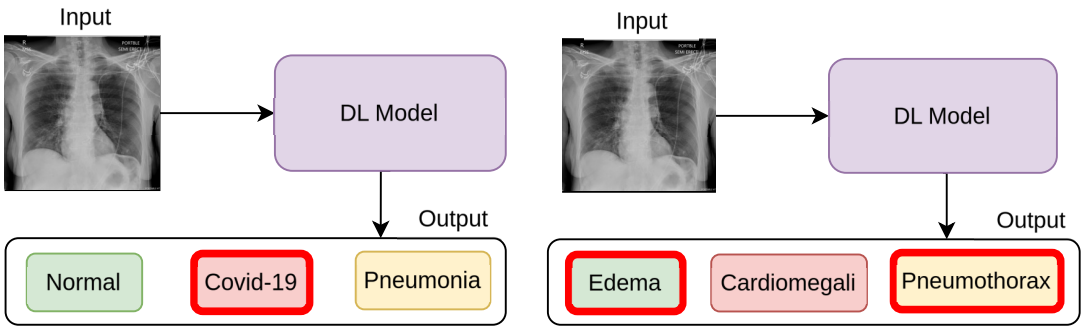
Multi-class classification (Fig. 3.2 (a)) constitutes a fundamental paradigm in supervised ML in which the objective is to assign each input instance $x \in \mathcal{X}$ to exactly one class label from a predefined set $\mathcal{Y} = \{y_1, y_2, ..., y_K\}$ where $K > 2$. This formulation generalizes the more constrained binary classification problem, which represents a special case where $K = 2$. The underlying discriminative function $f : \mathcal{X} \rightarrow \mathcal{Y}$ must effectively partition the feature space into K disjoint regions corresponding to each class. In Fig. 3.3 (a), we illustrate the decision boundaries that emerge in multi-class classification scenarios. DL models must learn to distinguish between classes based on input features, ensuring accurate predictions across the entire label space.

Several methodological approaches have been proposed to address multi-class scenarios. The one-versus-all (OVA) decomposition, also termed one-versus-rest, constructs binary classifiers K, each trained to discriminate between a particular class and the aggregation of all remaining classes [93]. Alternatively, the one-versus-one (OVO) decomposition employs $\frac{K(K-1)}{2}$ binary classifiers, each trained on pairs

of classes, with the final classification determined through voting mechanisms or other aggregation strategies [26]. Error-correcting output codes (ECOC) represent another approach, encoding each class with a unique binary string and training binary classifiers for each bit position [17].

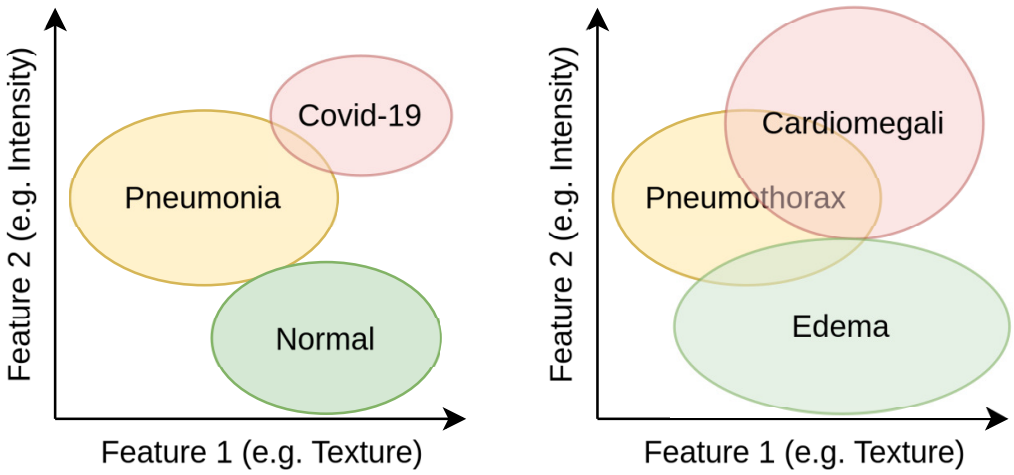
The difficulty of multi-class classification scales non-linearly with the cardinality of $\mathcal{Y}$. This complexity manifests in several dimensions:

- the curse of dimensionality becomes more pronounced;
- class imbalance effects are potentially exacerbated;
- decision boundaries become increasingly intricate, particularly in regions where multiple classes exhibit proximity or overlap in the feature space.



(a) Multi-class Classification (b) Multi-label Classification

Figure 3.2: Classification Paradigms in Medical Imaging.



(a) Multi-class Decision Boundaries (b) Multi-label Overlapping Regions

Figure 3.3: Feature Space Representation.

Additionally, as K increases, the discriminative capacity required of the model grows substantially, necessitating more sophisticated architectures and regularization techniques to mitigate overfitting.

In the context of medical imaging, multi-class classification might involve differentiating among various disease phenotypes, tissue types, or anatomical structures, each represented as a distinct class (Fig. 3.3 (a)). Contemporary approaches frequently leverage deep CNNs to learn hierarchical features directly from imaging data, obviating the need for hand-crafted feature engineering prevalent in traditional ML paradigms.

3.3 Multi-label Classification

Multi-label classification (Fig. 3.2 (b)) represents a distinct formalism in which each instance $x \in \mathcal{X}$ may be associated with multiple class labels simultaneously from a set of labels $\mathcal{L} = \{l_1, l_2, ..., l_q\}$. Formally, this is represented as a function $f : \mathcal{X} \rightarrow 2^{\mathcal{L}}$, mapping each instance to a subset of the power set of $\mathcal{L}$. This contrasts fundamentally with multi-class classification, which enforces mutual exclusivity among labels. In Fig. 3.3 (b), we illustrate the overlapping regions, which means multiple labels that can be associated with one input. The model must predict the presence or absence of each label, capturing the complex interrelationships and dependencies among labels.

The multi-label paradigm inherently captures the complexity present in numerous real-world scenarios where single-label annotations prove to be insufficient. From a methodological perspective, several approaches have been developed to address this challenge. Binary relevance transforms the problem into q independent binary classification tasks, one for each potential label [130]. Although computationally straightforward, this approach neglects potential label correlations. The classifier chains [88] extend the binary relevance by incorporating the dependencies of the labels through a sequential prediction process, where each classifier in the chain includes the predictions of the preceding classifiers as additional features.

Label powerset approaches reframe the problem by treating each unique combination of labels as a distinct class within a multi-class framework [107]. Although this explicitly models label correlations, it suffers from a combinatorial explosion as q increases, leading to sparse training data for many label combinations. Algorithm adaptation methods directly modify existing algorithms to accommodate multi-label data, while problem transformation methods convert multi-label problems into one or more single-label problems amenable to conventional classification algorithms.

The principal challenges in multi-label classification extend beyond those en-

countered in multi-class scenarios. The model must capture complex label co-occurrence patterns and conditional dependencies between labels, which may reflect underlying biological or physical relationships in the domain. Computational complexity scales dramatically with the cardinality of the label, and the evaluation metrics become more nuanced, with partial correctness requiring specialized performance measures such as Hamming loss, subset accuracy, and various variants of the F measure.

In medical imaging, multi-label classification is particularly relevant when multiple pathologies, anatomical features, or diagnostic criteria may co-exist within a single image or case. Contemporary deep learning approaches frequently employ architectures with multiple output nodes, each corresponding to a specific label, often with a sigmoid activation function replacing the softmax characteristic of multi-class problems, allowing for non-mutually exclusive predictions across the label space.

3.4 CoAtNet

A central drawback of CNNs lies in their limited ability to model long-range dependencies within the input data. Although CNNs excel at capturing local spatial features through hierarchical structures, they struggle to efficiently incorporate global contextual information. To address this issue, researchers have introduced ViTs, originally proposed for natural language processing tasks [110]. ViTs employ self-attention to represent global dependencies throughout the input sequence, offering a more holistic interpretation of the data. Such capabilities are particularly beneficial for applications that require an understanding of long-range relationships, including image classification [18] and object detection [10]. However, ViTs also pose challenges because of their lack of inductive biases, necessitating larger datasets and considerable computational resources. Since ViTs operate on the full input rather than localized subsets, they require substantial memory and processing power, limiting their suitability in resource-constrained environments.

Hybrid architectures such as CoAtNet have emerged to combine the strengths of CNNs and ViTs, as illustrated in Fig. 4.2. CoAtNet utilizes depth-wise convolution (Eq. 3.7), leveraging the MBConv block [95] to model local spatial relationships. A fixed kernel is employed to capture receptive field information:

$$\mathbf{y}_i = \sum_{j \in \mathcal{L}(i)} \mathbf{w}_{i-j} \circledast \mathbf{x}_j \tag{3.7}$$

In the above equation, $\circledast$ denotes convolution, $\mathbf{y}_i, \mathbf{x}_j \in \mathbb{R}^D$ represents output and

input at position i, $\mathcal{L}(i)$ specifies the local neighborhood, and $\mathbf{w}_{i-j}$ indicates the weights of the kernel. By incorporating CNNs as the architectural backbone, CoAtNet efficiently extracts local features while controlling computational complexity.

Simultaneously, CoAtNet employs self-attention to capture global dependencies and long-range interactions within the data. The self-attention calculation, shown in Eq. 3.8, determines spatial relationships by examining pairwise similarities $(\mathbf{x}_i, \mathbf{x}_j)$:

$$\mathbf{y}_i = \sum_{j \in \mathcal{G}} \frac{\exp(\mathbf{x}_i^T \mathbf{x}_j)}{\sum_{k \in \mathcal{G}} \exp(\mathbf{x}_i^T \mathbf{x}_k)} \mathbf{x}_j \tag{3.8}$$

Here, $\mathcal{G}$ represents the global spatial domain, while the denominator corresponds to the weight of the attention. CoAtNet integrates both self-attention and depth-wise convolution as formulated in Eq. 3.9:

$$\mathbf{y}_i^{pre} = \sum_{j \in \mathcal{G}} \frac{\exp(\mathbf{x}_i^T \mathbf{x}_j + \mathbf{w}_{i-j})}{\sum_{k \in \mathcal{G}} \exp(\mathbf{x}_i^T \mathbf{x}_k + \mathbf{w}_{i-k})} \mathbf{x}_j \tag{3.9}$$

The hybrid design of CoAtNet is particularly advantageous for Medical Image Analysis (MIA), where images often encompass complex structures and require both localized detail recognition and global contextual understanding. By efficiently combining both perspectives, CoAtNet can deliver accurate and reliable results. Moreover, its computational efficiency supports deployment in environments with limited resources and the need for near-real-time analysis. Despite its promise, CoAtNet remains underexplored for MIA applications.

4. Methodology

In this chapter, we discuss the general FL methodology pipeline of the FedPANs algorithm applied to CoAtNet with the application to multi-class and multi-label classification problems.

4.1 General Pipeline

Figure 4.1 provides a schematic overview of the FedPANs algorithms. In this framework, the FL server sets up a global model, which may be DenseNet121, ResNet50, MobileViT, or CoAtNet, with or without integrated PANs. These particular DL architectures have been selected for their ability to address the complexities of federated medical image classification tasks. DenseNet121, for example, leverages dense connectivity to facilitate feature reuse and help prevent vanishing gradients, thus ensuring both efficient computation and reliable performance. ResNet50's residual connections support deeper network training and have demonstrated proven effectiveness in medical imaging, making it an excellent choice for capturing subtle patterns within intricate datasets. Meanwhile, MobileViT combines attention mechanisms with mobile-optimized designs, enabling it to effectively handle local and global dependencies, an invaluable trait in resource-constrained FL settings. Finally, CoAtNet, as a hybrid model merging CNN-based local feature extraction with Transformer-driven global context modeling, offers a cutting-edge approach to manage both detailed and expansive image features.

Incorporation of PANs leverages the permutation invariance property of DL architectures. We introduce CoAtPENet, which integrates PANs into the CoAtNet model. Details of this integration are discussed in the following sections. After establishing the global model, it is distributed to participating clients. These clients receive data partitions organized either through Latent Dirichlet Allocation (LDA) [9] or a specialized custom splitting function designed for the specific problem at hand. Once clients complete their local training, the updated models are returned to the central server, where their parameters are aggregated to form an updated global

model. Repeating this process for N rounds progressively refines the global model. A summary of the general FL procedure is presented in Algorithm 1. In subsequent sections, we explore how FedPANs can be applied to both multi-class and multi-label classification scenarios.

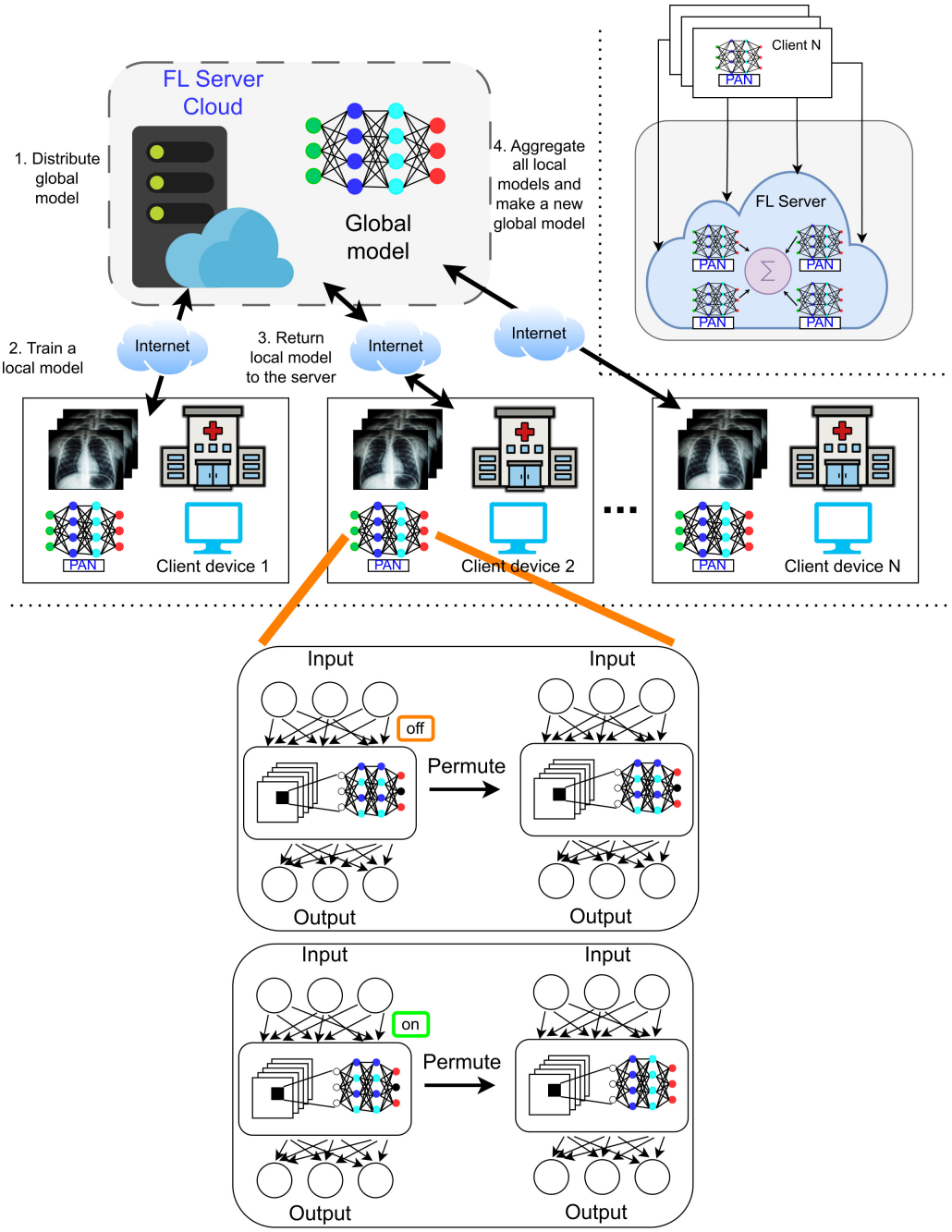


Figure 4.1: General Pipeline for Federated Medical Image Classification.

4.2 Non-iid and Imbalanced Data Creation

In this work, we distinguish between *non-iid* data for multi-class classification and *imbalanced* data for multi-label classification. Consequently, we present tailored data partitioning strategies that simulate realistic federated learning (FL) conditions for each classification type.

4.2.1 Non-IID Data for Multi-class Classification

A fundamental challenge in FL is that data distributions often differ significantly between clients, resulting in non-IID conditions. To emulate this heterogeneity, we employ Latent Dirichlet Allocation (LDA) [9], a method widely used in FL research to produce data splits reflecting the inherent diversity of the datasets of individual clients. Originally developed for topic modeling in natural language processing, LDA uncovers distributional patterns of topics and words. Translating this concept to federated image classification and following the guidelines in [38], we assume that each client independently draws training samples where class labels follow a categorical distribution across C classes, described by a probability vector $\mathbf{p}$. By sampling $\mathbf{p}$ from a Dirichlet distribution characterized by a concentration parameter α, we achieve varying degrees of data skew. This approach ensures that some clients receive a disproportionate share of specific classes, while others have distinctly different class compositions, effectively simulating non-IID scenarios.

To implement this, we define parameters that include the image dataset $\mathbf{X}$ with labels $\mathbf{Y}$, the number of partitions N, and the concentration parameter α governing the level of similarity or dissimilarity among clients. Our partitioning function

Algorithm 1 FL Algorithm (FedAvg, FedProx)

Input: Communication Rounds R, Number of Participants S, Dataset D, Batch Size B, Local Epochs E, and Learning Rate η
Output: Global Model w^G

1: Initialize: w_0^G
2: **for** $r \leftarrow 1$ to R **do**
3: Randomly distribute the dataset D and randomly choose a subset C_r of S participants from $\{1, 2, \ldots, C\}$
4: **for** each participant $j \in C_r$ in parallel **do**
5: $w_r^j \leftarrow \text{localTraining}[j](w_{r-1}^G, B, E, \eta)$
6: $\min_w h_k(w, w_r^j) = w + \frac{\mu}{2}\|w - w_r^j\|^2$
7: **end for**
8: $w_r^G \leftarrow \sum_{j=1}^{m} \frac{n^j}{N} w_r^j$
9: **end for**

includes safeguards to handle situations where the total number of samples does not divide perfectly among N clients. By doing so, it can generate either IID or non-IID partitions according to user-defined preferences. Using the LDA-based partitioning scheme, we create realistic and diverse data distributions, allowing a more thorough evaluation of model performance under heterogeneous FL conditions.

4.2.2 Imbalanced Data for Multi-label Classification

While LDA works effectively for single-label datasets, extending it to multi-label settings poses additional complications. Challenges arise in determining an appropriate label count per image, managing label inter-dependencies, and interpreting results meaningfully. Multiple labels per image complicate the process of assigning these images to different participants. As a result, careful consideration or alternative approaches are needed to generate suitable partitions for multi-label datasets.

Previous methods [51, 31] often partition multi-label datasets into equal-sized subsets, but this oversimplifies the complexity of FL scenarios. Equal divisions fail to capture the inherent imbalances present in real-world multi-label data distributions and disregard the likely heterogeneity among clients. This uniform split hinders the model's ability to learn from the diverse patterns that are typically encountered in practical FL environments.

To address this limitation, we propose a straightforward yet effective method to induce *imbalance* in multi-label data. Rather than relying on naive random selection, which can produce undesirable variance, we divide a dataset of D images into precisely N participants under well-defined constraints. We specify a minimum (m) and maximum (M) number of images per participant, ensuring that each client receives a sufficient amount of data to maintain reliable FL performance. This partitioning technique produces imbalanced distributions that more closely resemble real-world conditions, allowing the model to better adapt to the variability inherent in multi-label tasks. The pseudocode for this custom partitioning approach is presented in Algorithm 2. “‘

It should be noted that the data partitioning approach employed in this work is heavily based on custom partitioning under defined constraints. This strategy fosters a more representative and balanced distribution of data between clients, thereby reducing potential biases in model performance on previously unseen data. By encouraging a broader and more heterogeneous sample of the dataset among clients, this method effectively overcomes the drawbacks associated with simple random splitting.

Algorithm 2 Custom Data Splitting

Input: Given a dataset $\mathcal{S}$ of D images, number of participants N, minimum number of images m, maximum number of images M

Output: Split the dataset into N partitions, ensuring each corresponds to a participant and contains at least m images.

1: **for** $i = 1$ to N **do**
2: Select the number of images d_i within range $[m, M]$ for participant i
3: Randomly pick a subset $\mathcal{S}_i$ of d_i images from the set $\mathcal{S}$ for participant i
4: $\mathcal{S} = \mathcal{S} - \mathcal{S}_i$
5: **end for**
6: Distribute the remaining images (if any) equally to the participants

4.3 CoAtNet for Multi-class and Multi-label Classification

Multi-Class Classification. CoAtNet's deep hierarchical structure adeptly captures complex patterns within input data, making it highly suitable for multi-class classification challenges. We modify the final layers of CoAtNet to correspond to the intended number of output classes, as presented in Fig. 4.2. A softmax activation function in the final layer yields a probability distribution across the classes. The model training process employs cross-entropy loss, minimizing the divergence between predicted probabilities and true labels, thus ensuring that CoAtNet effectively assigns each sample to one of the available categories.

Multi-Label Classification. For multi-label classification, where samples can belong to multiple classes simultaneously, CoAtNet is adjusted to generate reliable predictions for each associated label. This involves replacing the final softmax activation with sigmoid functions, allowing each output neuron to produce an independent probability for its respective class (Fig. 4.2). The training procedure uses binary cross-entropy loss, guiding the model to manage complex label interrelationships and produce accurate multi-label predictions.

4.4 CoAtPENet

A primary objective of this study is to incorporate FedPANs into convolution-based and attention-based architectures, as shown in Fig. 4.3. We selected CoAtNet because of its balanced integration of convolutional and attention layers. Adopting a C-C-T-T configuration, where **C** indicates convolutional blocks and **T** transformer blocks, we apply additive and multiplicative PANs exclusively to the convolution blocks, as

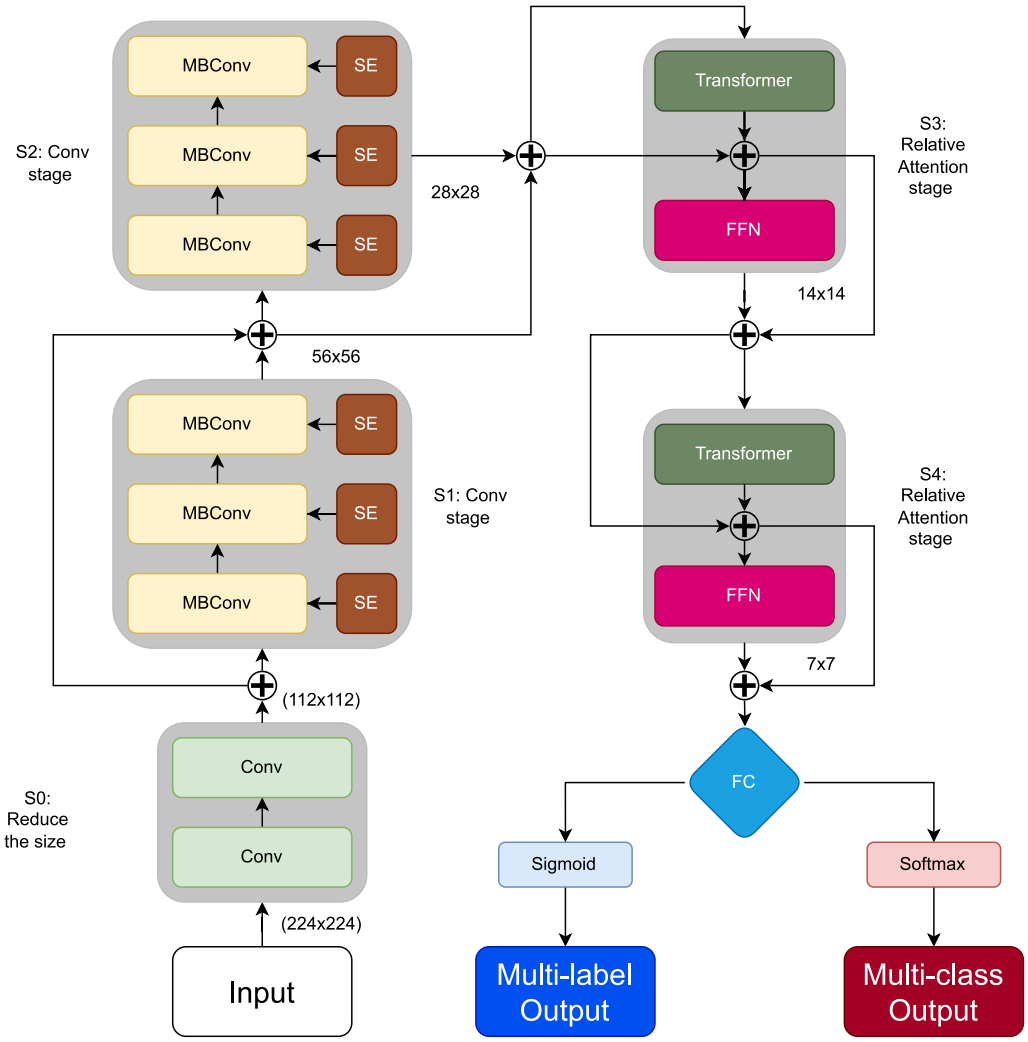


Figure 4.2: CoAtNet architecture with multi-class and multi-label outputs.

indicated by Eqs. 3.3 and 3.4:

$$\mathbf{PAN}_{+} : h_l = f_l((\mathbf{W}_l \mathbf{h}_{l-1} + \mathbf{b}_l) + \mathbf{e}_l) \tag{4.1}$$

$$\mathbf{PAN}_{\odot} : h_l = f_l((\mathbf{W}_l \mathbf{h}_{l-1} + \mathbf{b}_l) \odot \mathbf{e}_l) \tag{4.2}$$

In the convolutional stages (S1, S2) of Fig. 4.3, PANs is applied through internal weight permutation matrices, inducing stochastic transformations of feature representations without compromising discriminative capacity. In the upper-left part, we show how PANs is applied to the SE and MBConv blocks, which have critical components such as the point-wise and depth-wise convolutional kernels. The math notations shown in Eq. 4.1 and Eq. 4.2 for additive (+) and multiplicative ($\odot$) PANs are implemented by adding or multiplying the position encodings $\mathbf{e}_l$ to the neuron outputs, which was presented previously in Eqs. 3.3 and 3.4. h_l is the output on the

layer l, f_l denotes the activation function, and $l \in \{0, 1, \ldots, L\}$ is the layer index. The convolution parameters $\mathbf{W}_l \in \mathbb{R}^{C_l \times w_l \times h_l \times C_{l-1}}$ are defined by the input/output channels (C_l, C_{l-1}) and the dimensions of the kernel (w_l, h_l). The convolution layers are transformed according to $\mathbf{W}' = P_{\text{conv}}(\theta) \cdot \mathbf{W} \cdot P_{\text{conv}}^T(\theta)$, where the weights of the convolution layer $\mathbf{W}$ are updated according to the permutation matrix P_{conv}. The bias term is $\mathbf{b}_l$, and the position encoding $\mathbf{e}_l$ is generated using a sinusoidal function, as described in Eqs. 3.5 and 3.6 of Section 3.1.1.

Within the attention-based phases (S3, S4) shown in Fig. 4.3, the permutation mechanism, illustrated by green connectors, has been precisely extended to the Transformer and Feed-Forward Network (FFN) elements. We incorporate PANs into the attention blocks $\mathbf{T}$ (Eq. 3.8) of CoAtNet, as shown in Eqs. 4.3 and 4.4:

$$\mathbf{PAN}_{+} : h_l = f_l((\mathbf{A}_l \mathbf{h}_{l-1} + \mathbf{b}_l) + \mathbf{e}_l) \tag{4.3}$$

$$\mathbf{PAN}_{\odot} : h_l = f_l((\mathbf{A}_l \mathbf{h}_{l-1} + \mathbf{b}_l) \odot \mathbf{e}_l) \tag{4.4}$$

In these equations, $\mathbf{e}_l$ encodes position information for output neurons, f_l is the activation function, and h_l the layer-wise output with $l \in \{0, 1, \ldots, L\}$. The bias term is $\mathbf{b}_l$, and the self-attention parameters are $\mathbf{A}_l \in \mathbb{R}^{H \times W \times C}$, where H and W represent the spatial dimensions and C the number of channels. The position encoding $\mathbf{e}_l$ is generated using a sinusoidal function described in Eqs. 3.5 and 3.6 for both additive and multiplicative PANs. Visually, for the transformer-based components in the upper right panel, the permutation mechanisms P_{attn} have been adapted to accommodate the unique computational structures inherent in self-attention. Specifically, FedPANs operates on: (1) the query (Q), key (K), and value (V) projection matrices within the Multi-Head Self-Attention (MHSA) module, ensuring that the linear transformations applied to input representations; (2) the attention score patterns, transforming the attention matrix $\mathbf{A}$ according to $\mathbf{A}' = P_{\text{attn}}(\theta) \cdot \mathbf{A} \cdot P_{\text{attn}}^T(\theta)$, thereby preserving the relative importance distributions while obscuring the absolute relationship mappings; and (3) FFN weight matrices, which undergo permutation to protect the parameterization of the position-wise fully connected transformations.

Lastly, we examine how the application of PANs to the convolutional and attention layers influences the behavior of CoAtNet. In Fig. 4.3, the use of different colors (yellow for convolutional and green for attention permutations) highlights the domain-specific modifications necessary for successful integration. By integrating PANs first into the convolution layers and then into the transformer blocks, we achieve position-aware neurons throughout the entire architecture. The additive (Eq. 4.5) and

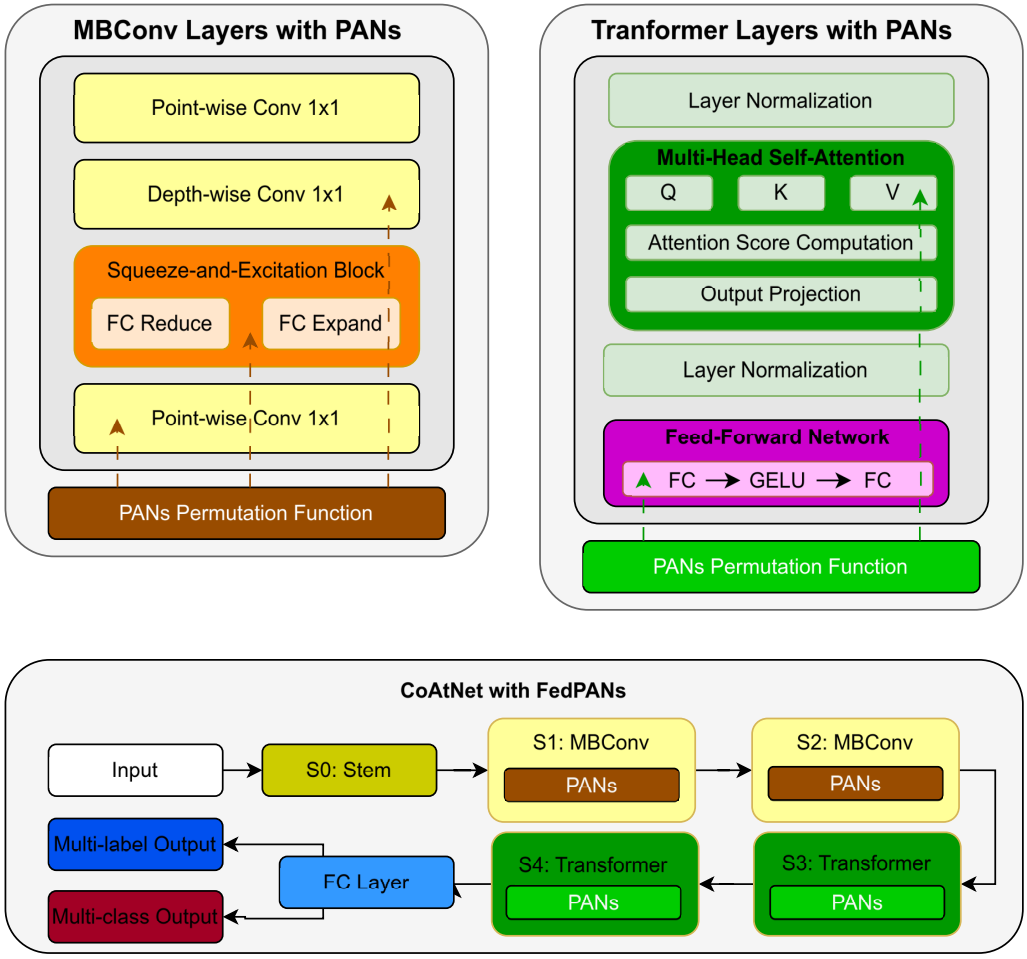


Figure 4.3: Detailed Layer Analysis of of CoAtNet with FedPANs.

multiplicative (Eq. 4.6) variants of PAN are defined as follows:

$$\mathbf{PAN}_{+} : h_l = f_l((\mathbf{W}_l \mathbf{h}_{l-1} + \mathbf{b}_l) + \mathbf{e}_l) + f_l((\mathbf{A}_l \mathbf{h}_{l-1} + \mathbf{b}_l) + \mathbf{e}_l) \tag{4.5}$$

$$\mathbf{PAN}_{\odot} : h_l = f_l((\mathbf{W}_l \mathbf{h}_{l-1} + \mathbf{b}_l) \odot \mathbf{e}_l) + f_l((\mathbf{A}_l \mathbf{h}_{l-1} + \mathbf{b}_l) \odot \mathbf{e}_l) \tag{4.6}$$

Here, $\mathbf{W}_l$ and $\mathbf{A}_l$ denote the convolutional and self-attention layers, respectively. By seamlessly incorporating PANs into CoAtNet, we enable the model to leverage position awareness at all layers, potentially enhancing its adaptability and performance in complex medical image analysis tasks.

5. Experimental Setup and Datasets

5.1 Experimental Settings

Our experimentation employs multiple deep learning architectures DenseNet121, ResNet50, MobileViT, and CoAtNet to address the complexities of our test scenarios. DenseNet121 is included due to its dense connectivity, which enhances feature reuse and alleviates the vanishing gradient issue. ResNet50, renowned for its residual connections, can accommodate deeper training networks without sacrificing stability. MobileViT, on the other hand, offers a balance between accuracy and computational efficiency, integrating attention mechanisms with mobile-optimized strategies to thrive in resource-constrained federated learning (FL) conditions. Finally, CoAtNet, a hybrid model that fuses convolutional and attention-based methodologies, is incorporated to illustrate the potential of such an integrated approach. Moreover, compared to other hybrid architectures ConViT [15], CvT [115], and UniFormer [54] our evaluation remains comprehensive. All models process images at a default resolution of 224x224, except MobileViT, which uses a 256x256 input size. For multi-class classification tasks, we use accuracy and the F1 score as performance indicators, while for multi-label classification we measure the mean AUROC score.

Regarding the PAN configuration, we introduce the parameters T and B as described in Eqs. 3.5 and 3.6 for additive and multiplicative PANs. Following the standard settings of [59], we select $T \in \{1.0, 4.0, 8.0\}$ and $B \in \{0.0, 0.05, 0.1, 0.15, 0.25, 0.5\}$. Empirically, the pairs chosen are: $(T, B) = (1.0, 0.05)$ for DenseNet121, $(4.0, 0.1)$ for ResNet50, $(1.0, 0.15)$ for MobileViT and $(4.0, 0.25)$ for CoAtPENet.

All experiments were executed on a DGX server equipped with five Tesla V100 SXM3 GPUs, each offering 32 GB of memory. The implementation relied on the PyTorch framework [5], the timm library [5], and the Flower framework [8] to simulate FL conditions. This setup ensured robust computational support for training and evaluation in federated scenarios.

5.1.1 Multi-class Classification

To analyze key hyperparameters in FL, we performed experiments using a subset of 20 participants drawn from the medium-sized CovidX dataset. In these trials, we designated 5 out of the 20 participants as active, set the total number of communication rounds to 100, and fixed the local training epochs at 3 for all runs. Our client-server framework is implemented using the Flower framework [8], enabling us to run experiments with different FL algorithms such as FedAvg, FedProx, and FedPANs. This setup allows flexible data partitioning methods both IID and non-IID across clients. We used LDA-based data splitting and varied the alpha parameter from 1 to 1000, where larger alpha values produce distributions closer to IID. The central server then aggregates the model weights reported by the clients according to the selected FL algorithm.

For the CovidX dataset, we fine-tuned DenseNet121, ResNet50, MobileViT, and CoAtNet models to address multi-class classification tasks. We adapted the classifier layer to match the number of classes required by FedAvg and FedProx. The classifier consists of a linear layer followed by a softmax activation function. During training, we applied the cross-entropy loss function to gauge the models' ability to learn discriminative features for multi-class classification. The experiments used the SGD optimizer with a learning rate of 1e-3 and applied batch sizes of 32 for DenseNet121 and ResNet50, and 64 for MobileViT, CoAtNet, ConViT, CvT and Uniformer models. The performance of the model was evaluated through the accuracy metric, calculated as the proportion of correctly classified images throughout the dataset.

5.1.2 Multi-label Classification

To thoroughly explore key FL hyperparameters, we expanded our experimental setup to include a total of 100 participants from large-scale datasets, as opposed to the smaller groups commonly examined in previous work. Additionally, we considered only a subset of these participants as active contributors to more closely approximate realistic FL scenarios. In actual deployments, not all participants, such as medical institutions or devices, can join every communication round due to constraints such as limited bandwidth, unstable connectivity, or restricted local resources.

In the case of the CheXpert and MIMIC-CXR datasets, we engaged 10 out of the 100 clients simultaneously. This fraction was empirically selected to ensure a manageable balance between diversity, training efficiency, and computational feasibility. We conducted 500 communication rounds in total and fixed the number of local epochs at 3 for all experiments. Running exactly three local epochs strikes

a practical equilibrium: it curtails the risk of overfitting when data are non-IID, accelerates training progress, and proves vital in resource-limited environments. This configuration reflects real-world FL conditions, where clients have limited computational capabilities and may be intermittently available. Limiting local epochs also reduces communication overhead and prevents scenarios in which slower clients become performance bottlenecks, thus facilitating steady and efficient global model refinement.

For our experiments, we used the AdamW optimizer, BCELoss as loss function, and a batch size of 32, maintaining a constant learning rate of 1e-3. To evaluate the performance of our models on multi-label classification tasks involving 14 labels, we relied on the mean AUROC metric. This measure evaluates the model's average discriminative ability across multiple labels, effectively capturing how well it distinguishes positive from negative instances for each label. Higher mean AUROC values indicate more effective handling of complex, multi-label scenarios.

5.2 Datasets

For the centralized training portion of our experiments, we divided all datasets into training, validation, and testing subsets. In contrast, for federated learning (FL) training, we employed two distinct data partitioning strategies: the LDA-based method and a custom partitioning function. To improve model robustness, we applied a variety of data enhancement techniques, including random horizontal flips, cropping, translations, scaling, and shearing. In addition, all training images were resized and normalized using ImageNet's mean and standard deviation parameters.

5.2.1 CovidX Dataset

To address multi-class classification tasks, we utilized the CovidX dataset [113], which consists of approximately 27,000 X-ray images categorized into three classes. We split the training portion using the LDA method, distributing the data among K clients. The label distribution of each client followed a Dirichlet distribution (α), managed through the Flower framework. In our experiments, we set $\alpha = 1/1000$. Figure 5.1 illustrates the resulting non-IID label distributions for the CovidX dataset.

5.2.2 CheXpert and MIMIC-CXR Datasets

For the multi-label classification experiments, we selected CheXpert [41] and MIMIC-CXR [45] datasets. CheXpert, one of the largest publicly available medical imaging

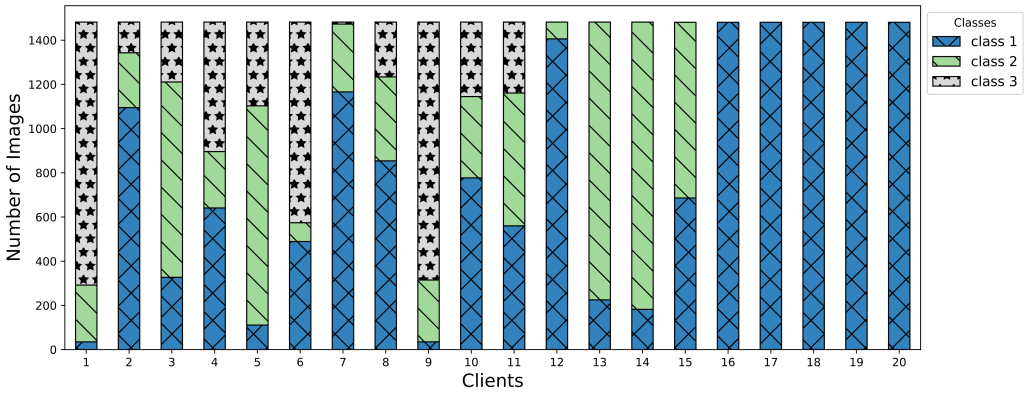


Figure 5.1: Non-IID data distribution of 20 clients for CovidX dataset using LDA partitioning.

datasets, contains 224,316 X-ray images from 65,240 patients covering 14 pathologies. Similarly, MIMIC-CXR is a large-scale dataset with 14 comparable pathologies. Both datasets include labels that can be 0 (absent), 1 (present), or -1 (uncertain). We addressed the uncertain (-1) labels employing the U-Ones approach [41], converting all -1 labels to 1.

For data splitting in these multi-label contexts, we implemented our custom partitioning function to distribute the data across 100 clients, inducing an imbalanced setup as depicted in Fig. 5.2. It is important to note that the figure does not strictly represent the exact number of images per client since, in multi-label scenarios, a single image can be associated with multiple labels. Consequently, the image count per client may increase, reflecting the complexity of label assignments within these datasets.

5.3 Ablation study on PANs

In section, we analyze the PANs hyperparameters selected for the DenseNet, ResNet, and MobileViT algorithms. In Fig. 5.3, DenseNet shows highly fluctuated results in almost all cases. For instance, even though we could reach the highest accuracies around rounds 50 to 60, we can observe the declining pattern in round 70. However, in later rounds, we can see that performances recovered and we decided to select $T = 1$ and $B = 0.15$ with additive $\mathbf{PAN}_{+}$. In Figs. 5.4 and 5.5, ResNet and MobileViT show better learning patterns compared to DenseNet. For example, increasing the value of T from 1 to 4 for the multiplicative $\mathbf{PAN}_{\odot}$ shows the best accuracy. For the additive $\mathbf{PAN}_{+}$, the settings $T = 1$ and $B = 0.0$ showed the best performance. Changing the values for T and B diminished the performance of ResNet with additive $\mathbf{PAN}_{+}$. MobileViT's performance looks the least fluctuated.

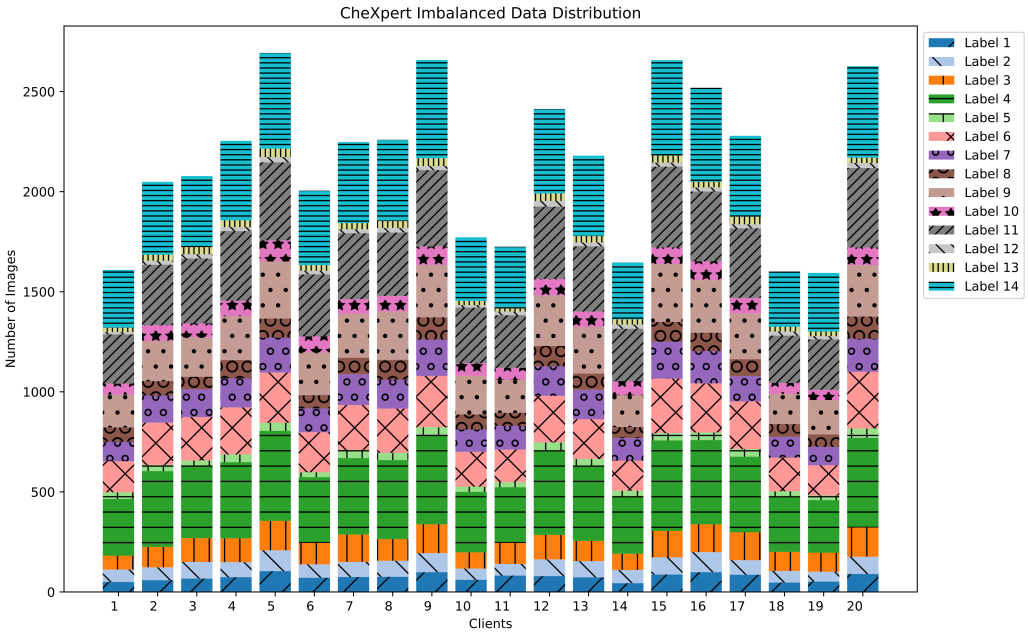


Figure 5.2: Stacked Imbalanced data distribution of first 20/100 clients for CheXpert dataset using custom partitioning.

We suspect that it is because of the attention mechanism which concentrates on the global view. However, due to data shortage, it could not outperform the ResNet model.

As an ablation study, we investigate the contribution of the different modules of the CoAtPENet model. As the CoAtPENet model is a hybrid model consisting of both convolution and attention modules, we assess the impact of each component on overall performance by progressively removing them and evaluating the performance on the test set. In general, we selected several period (T) and amplitude (B) values of the position encoding for additive $\mathbf{PAN}_{+}$ and multiplicative $\mathbf{PAN}_{\odot}$. In Fig. 5.6, we turn off the application of PANs in the attention layers. We can observe a gradual increase in accuracy, but the fluctuation is high. In Fig. 5.7, we applied PANs on the attention mechanism only. It also shows similar patterns as convolution with a high fluctuation rate. In both figures, it is hard to distinguish which parameter fits well. However, when PANs applied to both the convolution and the attention layers in Fig. 5.8, we can observe steady learning patterns. For example, all multiplicative cases $\mathbf{PAN}_{\odot}$ show a steady learning curve and get the highest accuracies. Thus, we decided to choose the $T = 4$ and $B = 0.1$ for the CoAtPENet architecture with the application of multiplicative $\mathbf{PAN}_{\odot}$ for both the convolution and attention mechanisms. The additive $\mathbf{PAN}_{+}$ shows a negative impact when increasing the parameters T and B. Setting higher parameters for T and B yields a higher fluctuation for $\mathbf{PAN}_{+}$.

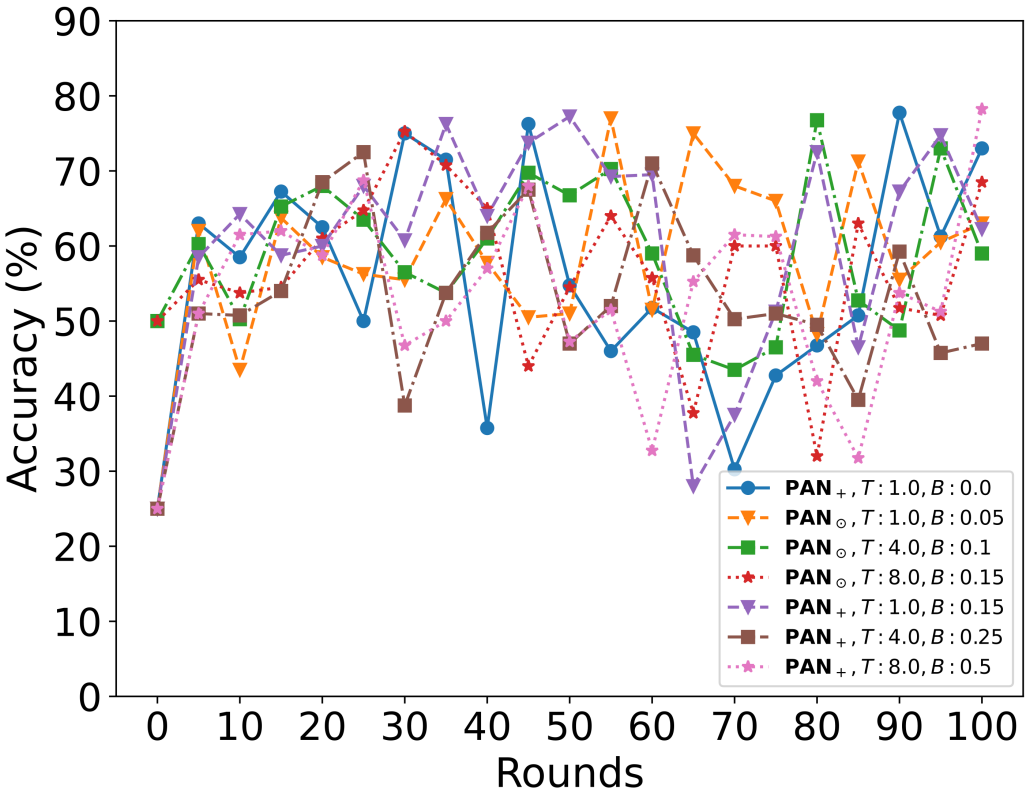


Figure 5.3: DenseNet with PANs on convolution modules.

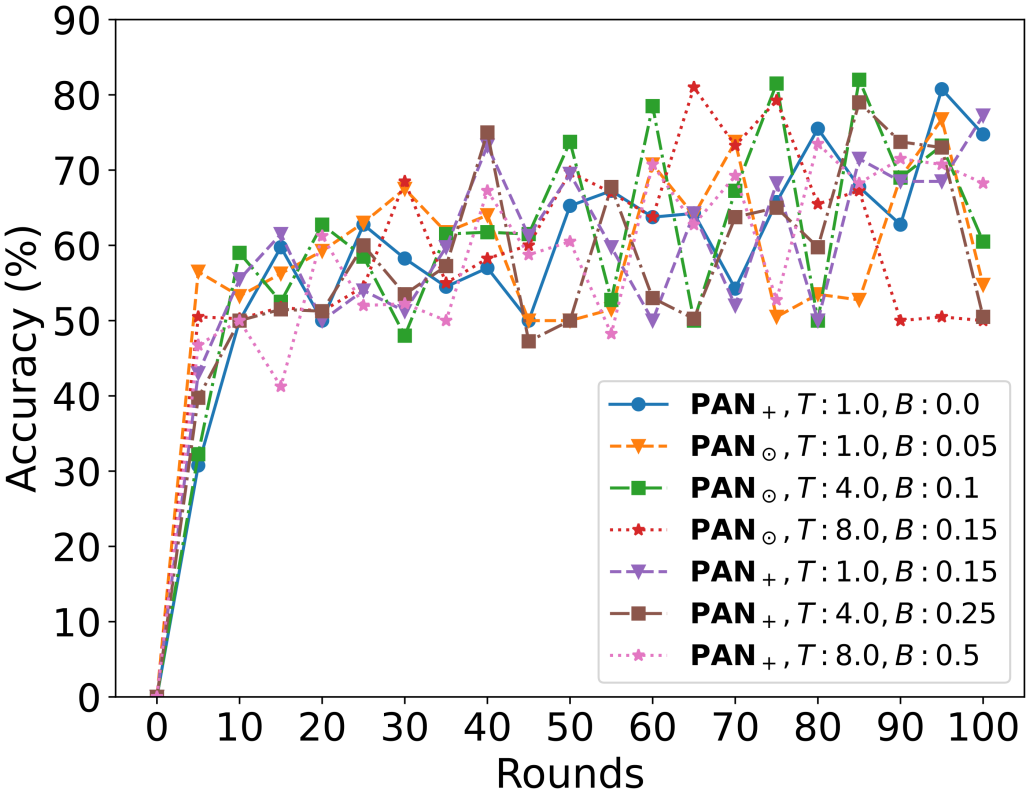


Figure 5.4: ResNet with PANs on convolution modules.

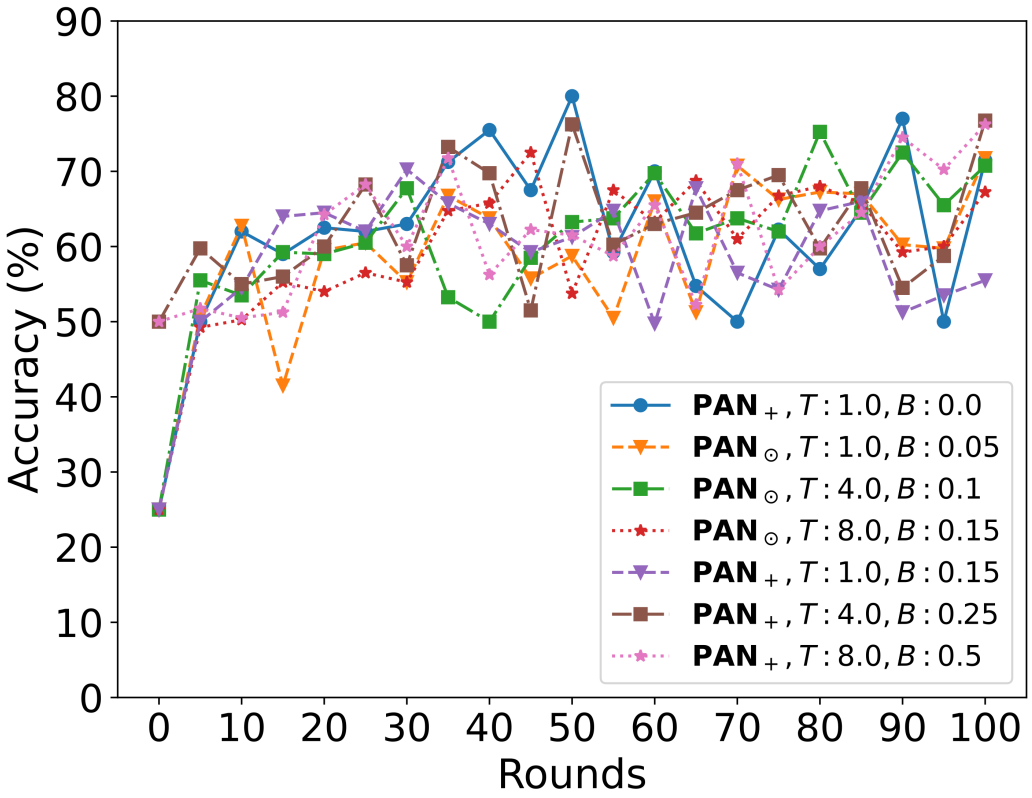


Figure 5.5: MobileViT with PANs on attention modules.

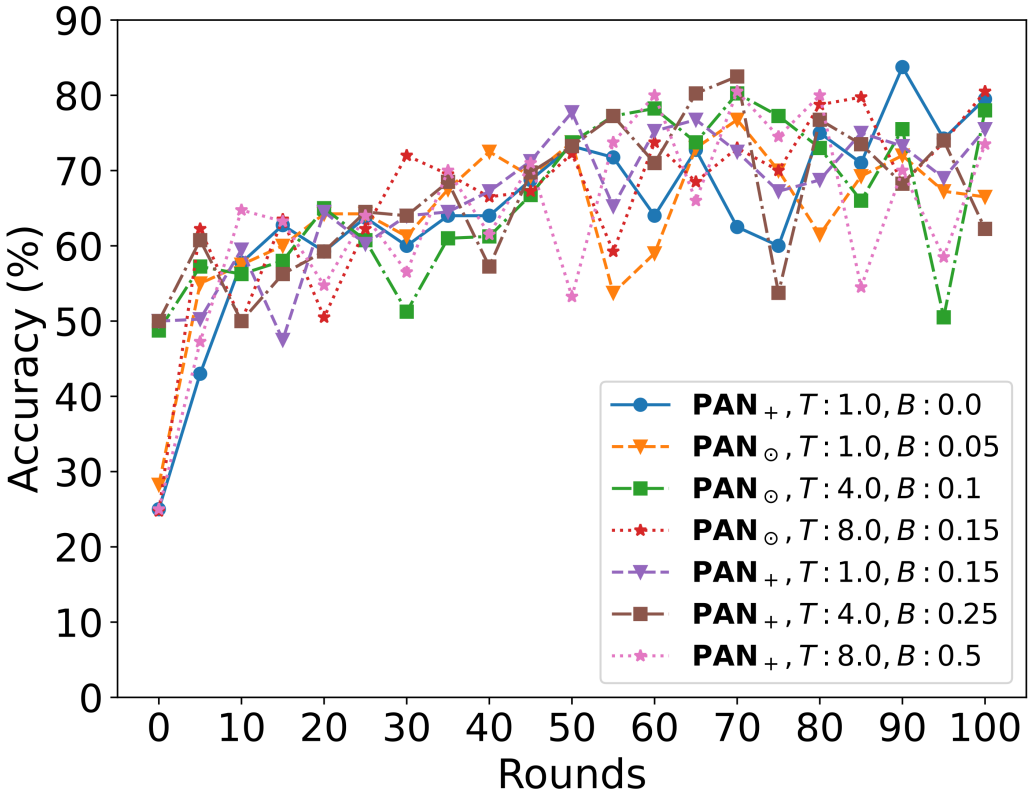


Figure 5.6: CoAtPENet convolution modules.

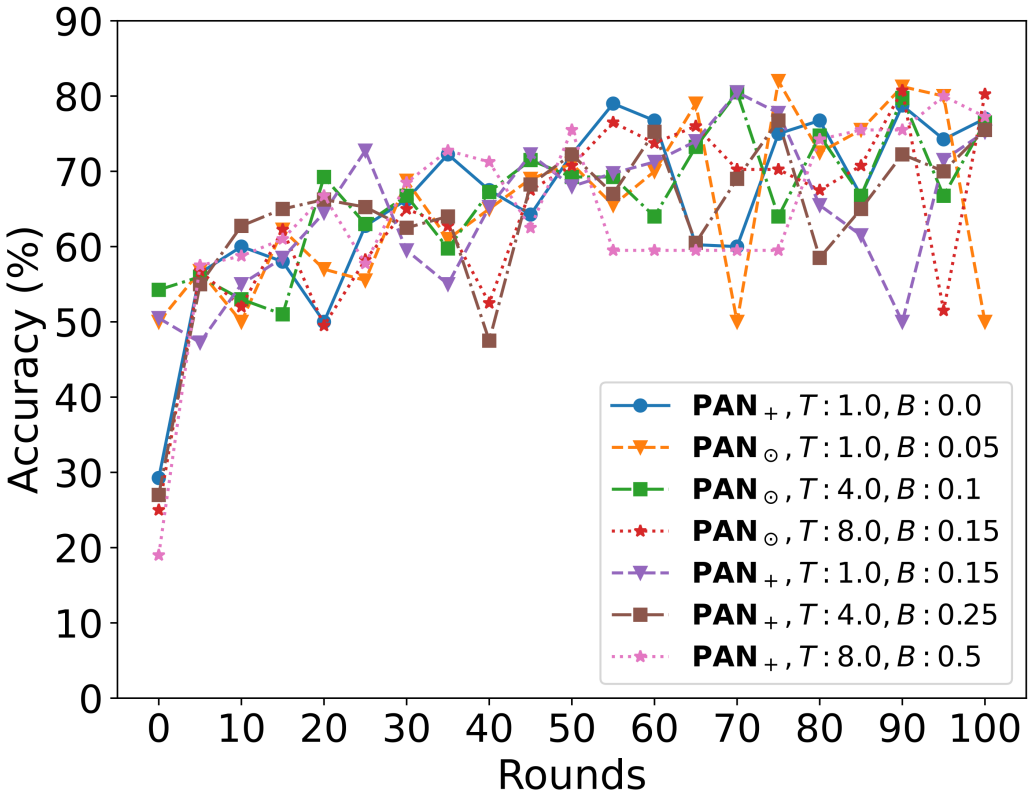


Figure 5.7: CoAtPENet attention modules.

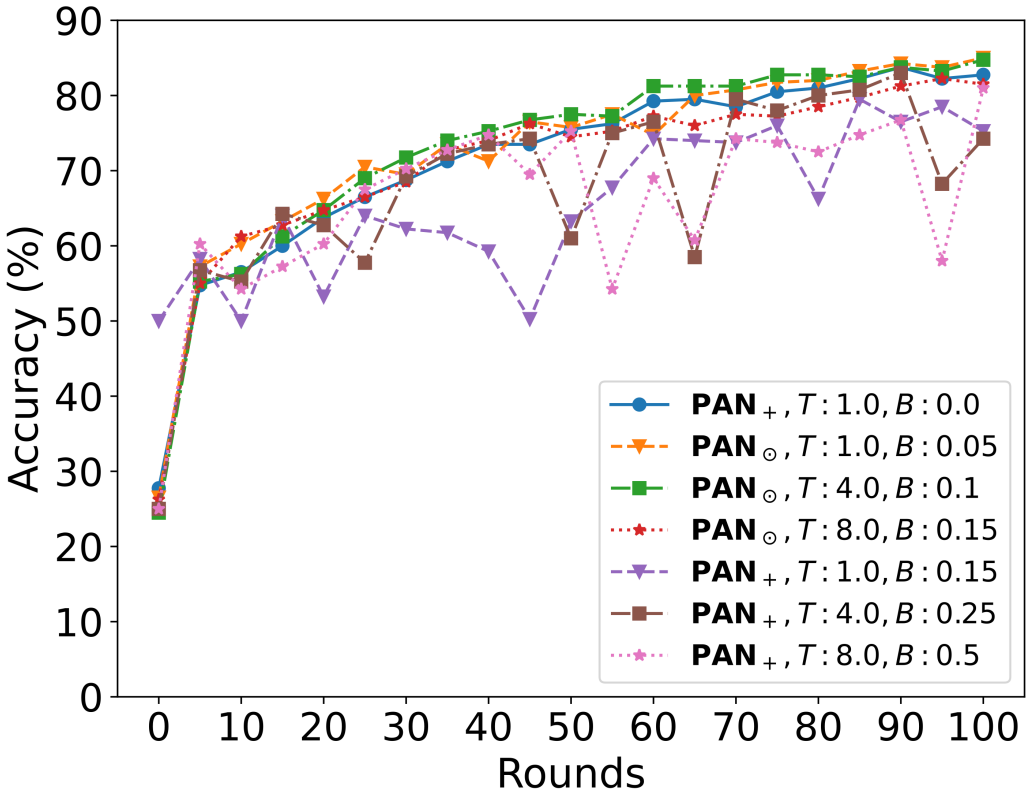


Figure 5.8: CoAtPENet combined attention and convolution modules.

6. Experimental Results

In this chapter, we discuss the experimental findings for various classification tasks with centralized and federated setups.

6.1 Performance Evaluation for Different DL Models

6.1.1 CovidX Dataset

First, we start by evaluating the performance of DL models for the multi-class classification. In Table 6.1, Figures 6.1 and 6.2, we present the results with and without PANs for seven different DL models. The centralized results are given in Table 6.1 (a), where the pre-trained models show an accuracy of more than 95% and an F1 score of more than 0.9. We can observe from Fig. fig:table1.1 that CNN-based pre-trained models show superior performance compared to other models. This is due to strong inductive bias and efficient feature extraction when the data is limited and knowledge is transferred. Unlike attention-based architectures that often require large datasets to learn effectively due to their minimal inductive biases, CNNs leverage hierarchical feature learning, enabling them to generalize better on smaller datasets. The pattern is seen clear for the IID and non-IID cases where pre-trained CNN models show higher accuracy and F1 score compared to attention-based and hybrid models. However, the overall performance of the pre-trained models is higher than 88% and with a reliable F1 score of more than 0.82. ConViT model is the worst performing model in all cases. This might be due to ConViT being heavily relying on the attention mechanism and the lack of strong inductive bias. Another reason might be that ConViT may overfit or fail to generalize due to insufficient data.

In contrast, untrained models show higher variability compared to pre-trained models in Fig. 6.2. For example, centralized models show more than 92% accuracy with a high F1 score except ResNet and ConViT. However, we can observe a sudden drop in accuracy for FL setup with IID and even lower performance in non-IID cases

Table 6.1: CovidX dataset results on FedAvg algorithm.

Models	Pre-train	Centralized		IID		non-IID	
		Acc. (%)	F1 score	Acc. (%)	F1 score	Acc. (%)	F1 score
Densenet121		95	0.95	84	0.65	70	0.60
Densenet121	✓	95.75	0.95	95	0.94	94	0.93
Resnet50		82	0.90	73	0.65	71	0.63
Resnet50	✓	97.75	0.96	96	0.95	95	0.94
ViT		78	0.72	63.5	0.59	60	0.56
ViT	✓	88	0.85	82.5	0.82	81.75	0.81
MobileViT		92	0.9	90	0.83	82	0.81
MobileViT	✓	97	0.95	94	0.85	92	0.89
CoAtNet		93	0.84	92	0.86	85	0.83
CoAtNet	✓	96.50	0.91	92	0.89	89	0.88
ConViT		58	0.40	57	0.55	55	0.41
ConViT	✓	62	0.55	55	0.53	53	0.52
CvT		92.5	0.92	73	0.71	69	0.66
CvT	✓	96.25	0.96	87	0.86	84	0.82
Uniformer		94	0.93	88	0.85	83	0.81
Uniformer	✓	96.7	0.96	90	0.89	88	0.88

(a) Results of different DL models without PAN.

Models	IID		non-IID	
	Acc. (%)	F1 score	Acc (%)	F1 score
Densenet	77	0.63	66	0.61
Resnet	86	0.75	77.25	0.73
MobileVit	89	0.87	90	0.85
CoAtPENet	92	0.89	92	0.87

(b) Results of different DL models with PAN on FedAvg.

for CNN models. This is due to the use of fewer clients, which means the use of less data compared to the centralized performance. MobileViT and CoAtNet models show a close performance to centralized accuracy in IID cases. All models struggle with non-IID cases showing an 8-15% drop compared to centralized accuracies. In the non-IID scenario, the highest accuracy is obtained by the CoAtNet model compared to other models. We include ViT results in Table **??** (a), but this model is not used in subsequent experiments due to its poor performance and unstable results with high fluctuation rates. Instead, we selected MobileViT as our attention-based model for all remaining experiments. Our findings suggest that the ViT architecture is not well suited for medical image classification tasks with limited data and non-IID scenarios.

Figure 6.4 compares the performance of four pre-trained models such as DenseNet121, ResNet50, MobileViT, and CoAtNet on the CovidX dataset under IID and non-IID FL settings. All models perform well in IID conditions, with CoAtNet and MobileViT

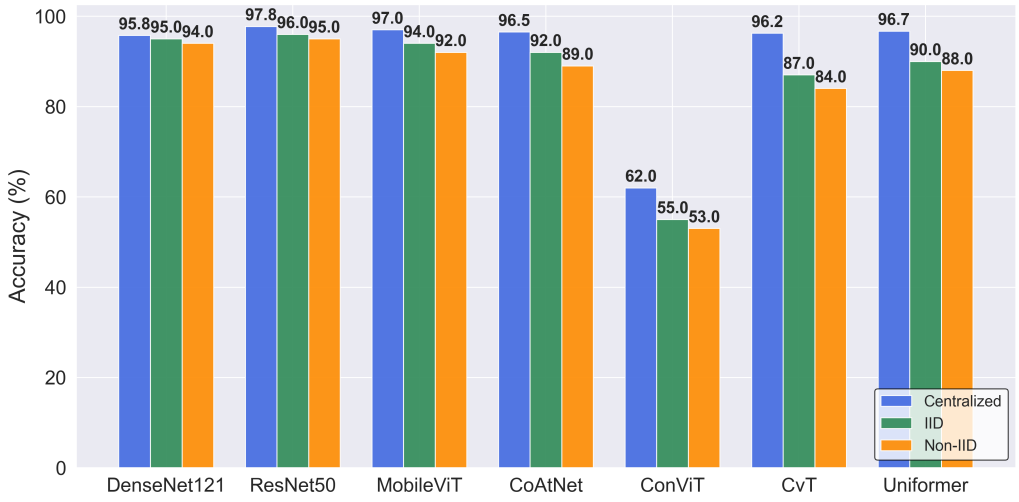


Figure 6.1: CovidX results of different DL models with pre-training.

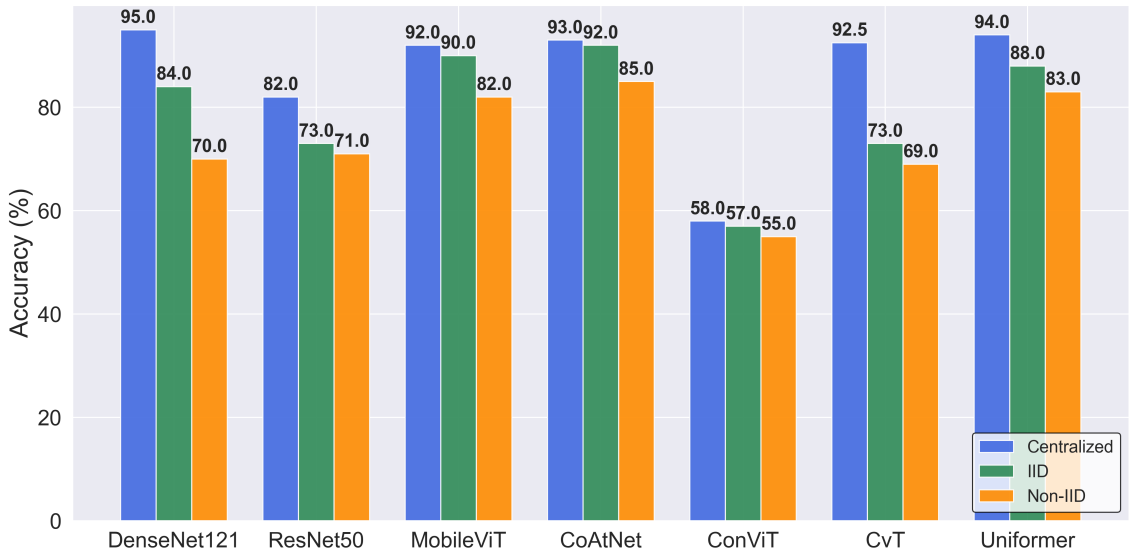


Figure 6.2: CovidX results of different DL models without pre-training.

showing smooth learning curves. In non-IID settings, CoAtNet outperforms others, maintaining consistent accuracy around 85%-90% with minimal fluctuations, demonstrating its robustness to heterogeneous data. MobileViT follows closely, exhibiting stable performance with fewer oscillations compared to DenseNet121 and ResNet50. DenseNet121 struggles the most in non-IID conditions, with significant accuracy drops and variability, while ResNet50 shows moderate improvement but remains less stable. Overall, CoAtNet and MobileViT stand out as the most adaptable architectures for FL in both IID and non-IID settings.

Figure 6.5 shows the FL performance of four DenseNet121, ResNet50, MobileViT, and CoAtNet models trained without pre-training under IID and non-IID conditions.

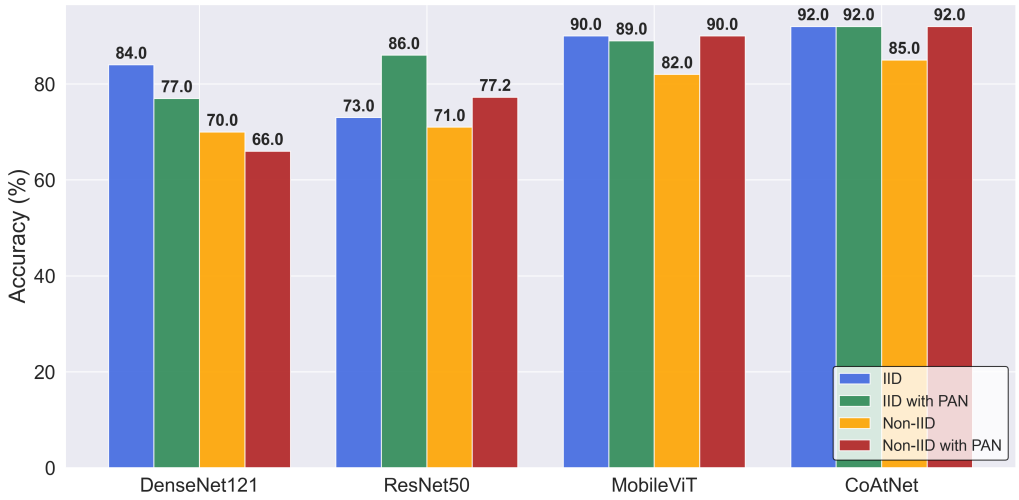


Figure 6.3: CovidX results of different DL models with and without PANs on FedAvg.

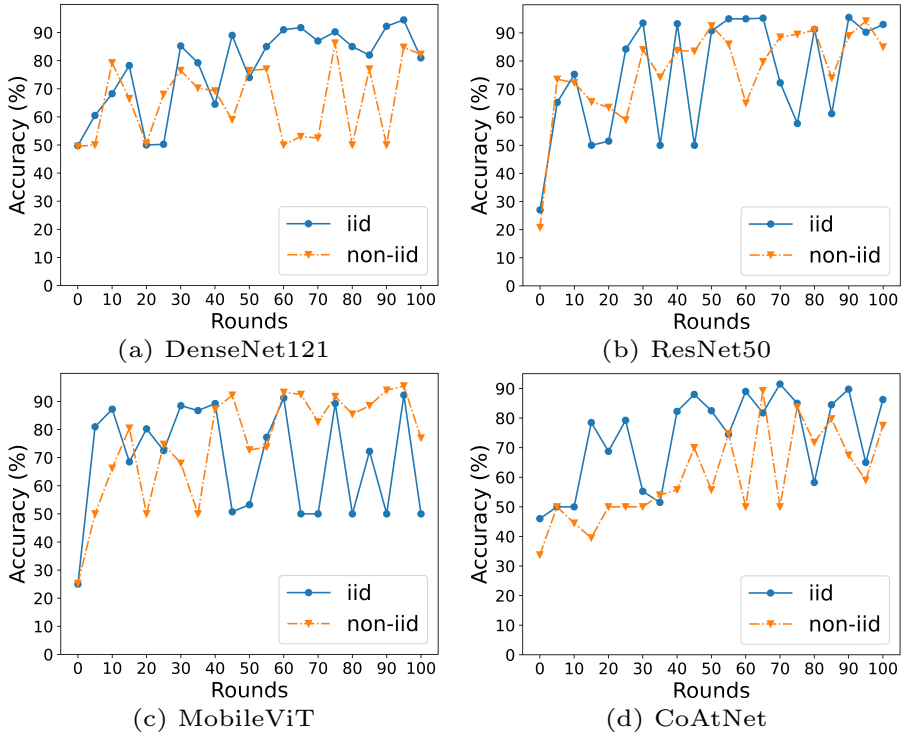


Figure 6.4: FedAvg results of 4 pre-trained models and different data distribution on CovidX dataset.

CoAtNet outperforms the others, achieving the highest accuracy in IID settings and maintaining strong performance under non-IID conditions, highlighting its robust hybrid architecture. MobileViT follows, showing better stability than DenseNet121 and ResNet50, especially in non-IID settings, where it reaches 50%-70% accuracy. DenseNet121 and ResNet50 exhibit significant performance drops and instability in non-IID scenarios, reflecting their sensitivity to data heterogeneity.

We observed a positive impact when we applied PANs. In Table 6.1 (b) and Figure 6.3, for all models in FL IID cases, the F1 score is improved and the accuracy of the ResNet model improved by 16% where other models' performance stayed the same. For non-IID cases the impact of PANs is significant that we can observe an improvement in performance and an F1 score for all models except DenseNet. It shows that PANs are the most useful in challenging scenarios.

In Figure 6.6, we can observe the performance of DL models with and without PANs in non-IID FL settings. Across all models, PANs significantly improve accuracy and stability. CoAtPENet benefits the most, achieving smoother and higher accuracy, compared to lower and more fluctuating performance without PANs. MobileViT also shows notable improvements, maintaining stable accuracy with PANs compared to fluctuating results without. DenseNet121 and ResNet50 see moderate gains, with PANs reducing variability and improving accuracy by around 10%-15%.

Let us take a look at Figure 6.7(a) which illustrates the non-IID FL performance of DenseNet, ResNet, and MobileViT with applied PANs. MobileViT demonstrates the highest accuracy and stability, consistently reaching 85% by the 50th round and maintaining it throughout training, highlighting its robustness in non-IID conditions. ResNet follows with moderate performance, achieving around 80% accuracy, but

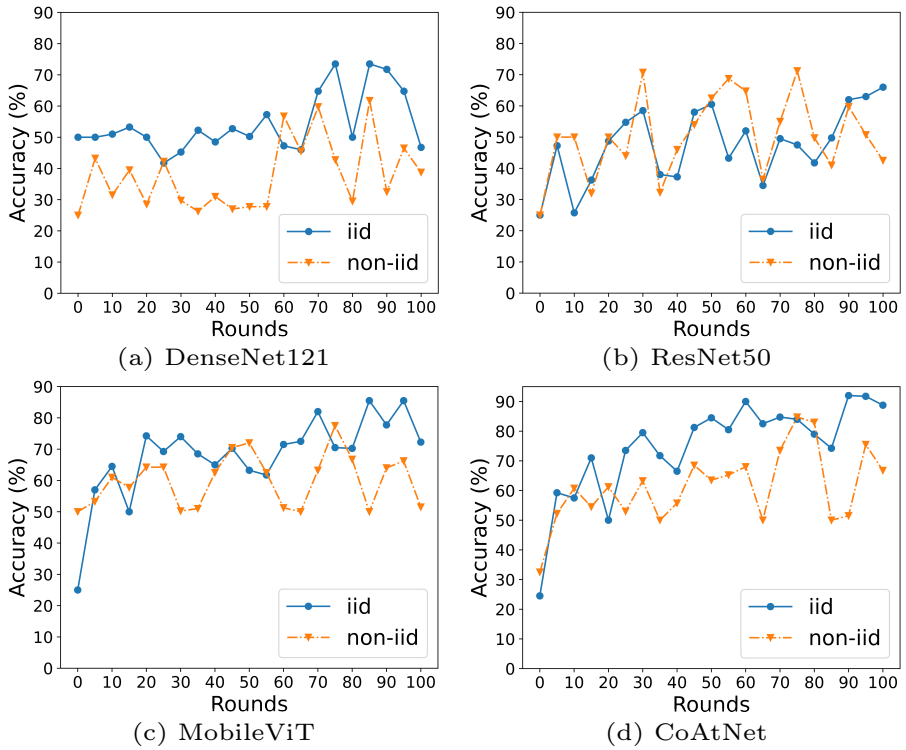


Figure 6.5: FedAvg results of 4 models with no pre-training and different data distribution on CovidX dataset.

exhibits more variability between rounds. DenseNet performs the worst, with significant fluctuations and lower overall accuracy, indicating its limited adaptability to non-IID FL environments even with PANs. These results emphasize MobileViT's strength in combining efficiency and robustness under data heterogeneity when the attention mechanism is enhanced with PANs.

Figures 6.7(b) and 6.7(c) illustrate the performance of the three architectural variants of CoatPENet: convolution-based (conv), attention-based (attn) and hybrid (conv + attn) in FL settings with non-IID, evaluated using FedAvg and FedProx algorithms. In both figures, the hybrid architecture (conv+attn) consistently achieves the best performance, leveraging its ability to combine the strengths of convolutional

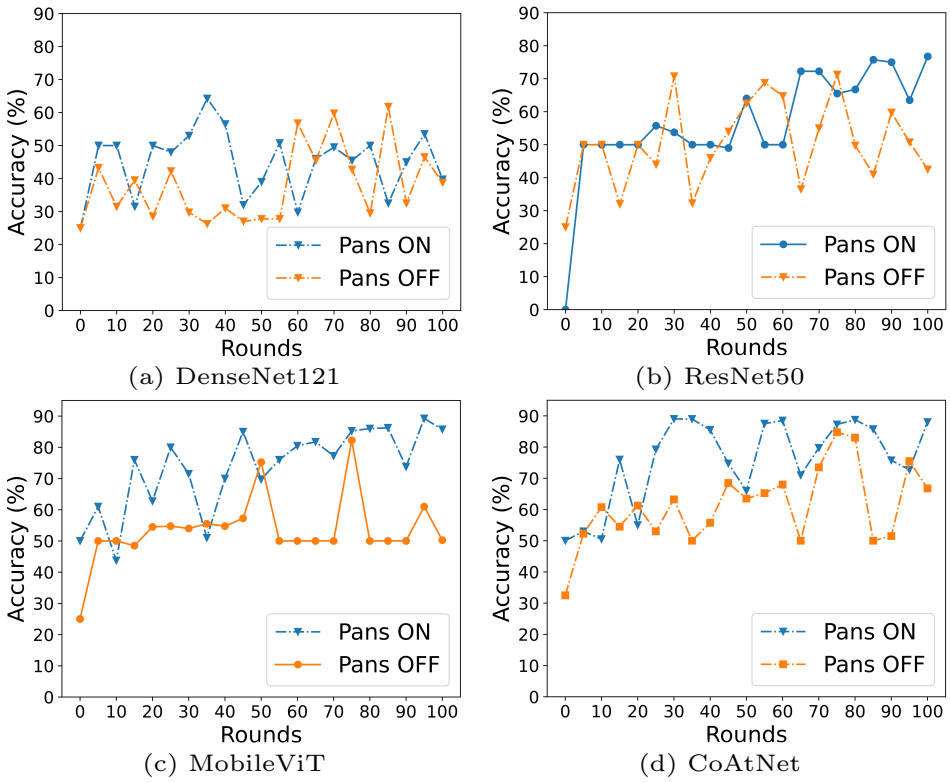


Figure 6.6: FL results of 4 DL models with PANs on/off on CovidX dataset.

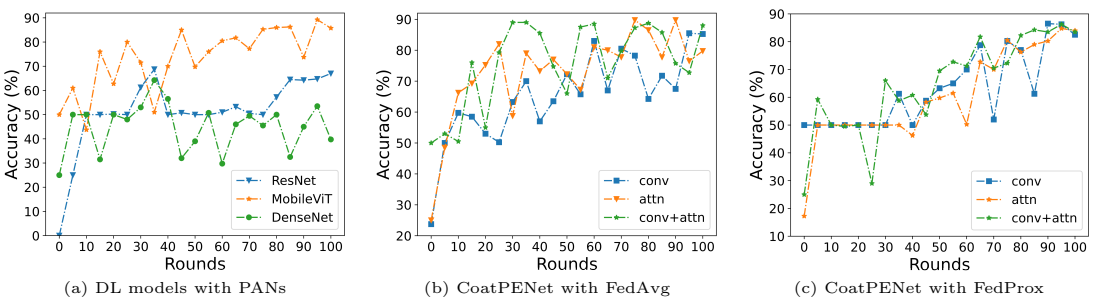
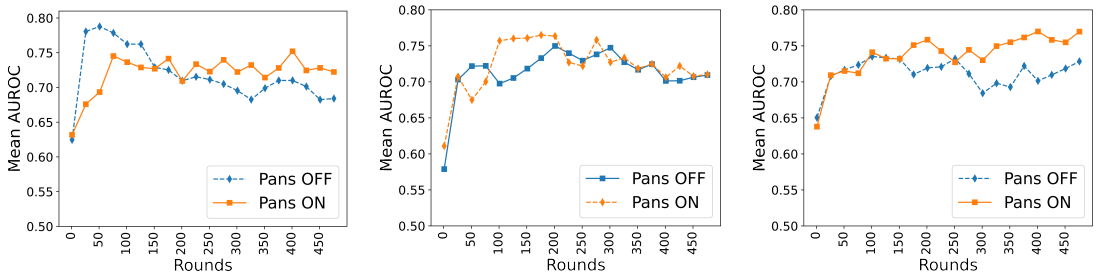


Figure 6.7: Non-iid FL results of different DL models with PANs on CovidX dataset.

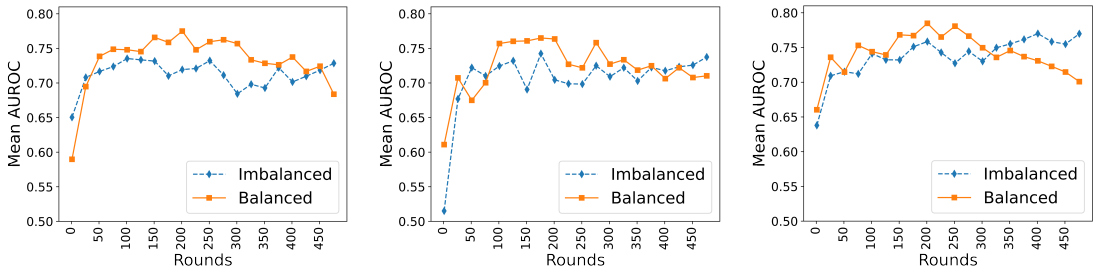


(a) DenseNet121 (b) MobileViT (c) CoAtNet

Figure 6.8: Imbalanced FedAvg results of DenseNet121, MobileViT, CoAtNet on CheXpert dataset.

layers for local feature extraction and attention mechanisms for capturing global dependencies. Under FedAvg, the hybrid model demonstrates robust accuracy, consistently reaching 85%-90% after 40 rounds and maintaining smooth convergence. The only attention-based model performs competitively, achieving 80%-85% accuracy, but with slightly greater variability, indicating a strong representation of global features but less adaptability to noisy heterogeneous data distributions. However, the only convolution-based model struggles the most, peaking at 75% accuracy and exhibiting significant fluctuations, reflecting its limited ability to handle non-IID distributions effectively.

When using the FedProx algorithm, performance trends remain consistent, but all models exhibit improved stability compared to FedAvg, with reduced instabilities. The hybrid model again outperforms the others, achieving accuracy levels of 88%-90% with smoother convergence, demonstrating its resilience in non-IID scenarios. The attention-based model also benefits from FedProx's stabilization mechanisms, maintaining an accuracy of 85%, albeit with minor fluctuations. The convolution-based model continues to underperform, reaching 75% accuracy, but with better consistency compared to its results with FedAvg. These findings emphasize that, while attention-based and hybrid models are better suited for non-IID FL environments, the hybrid architecture provides the best balance between performance and stability. Moreover, FedProx proves to be an effective algorithm for improving convergence and reducing variability, especially for architectures such as the convolution-based model, which struggle under more traditional algorithms like FedAvg. In general, the results highlight the importance of integrating advanced hybrid architectures with optimized FL algorithms to address the challenges of data heterogeneity in federated learning.



(a) DenseNet121 (b) MobileViT (c) CoAtPENet

Figure 6.9: FedAvg results of DenseNet121, MobileViT, CoAtNet on CheXpert dataset.

6.1.2 CheXpert and MIMIC-CXR Datasets

In this section, we discuss the performance of different DL models for multi-label classification task on two datasets. Figure 6.8 presents how the mean AUROC scores evolve over multiple communication rounds for DenseNet121, MobileViT, and CoAtNet. Incorporating PANs improves the convergence of all models in a federated context, especially under conditions where the data distribution is imbalanced. Referring to Figure 6.8(a), DenseNet121 without PANs initially achieves the highest mean AUROC score during the early communication rounds, after approximately the 50th round, the PAN-enhanced variant overtakes its counterpart without PANs. This shift underscores the long-term advantages of PANs in improving the robustness and stability of the model over time.

In contrast, MobileViT with PANs, depicted in Figure 6.8 (b), shows minimal performance improvement compared to the scenario in which PANs are disabled. This model exhibits greater performance fluctuations, highlighting a more unstable training process. The volatility suggests that while PANs may not significantly enhance MobileViT's performance, they do contribute to the model's adaptability in federated learning environments. CoAtPENet, illustrated in Figure 6.8(c), exhibits a similar performance trend when PANs are applied. The performance gap between the ON and OFF conditions of the PAN tends to narrow as the number of communication rounds increases, suggesting that the benefits of the PANs become more pronounced with extended training. In general, these findings underscore the importance of incorporating PANs in FL frameworks to enhance model performance, particularly in challenging scenarios characterized by data imbalance.

Figure 6.9 illustrates the FedAvg performance of the DenseNet121, MobileViT, and CoAtPENet models in the CheXpert dataset, showing notable results in balanced and unbalanced conditions. The analysis reveals that all three models benefit significantly from PANs, which improve model performance under imbalanced conditions

by approximately 2- 4% in the AUROC scores. This improvement highlights the effectiveness of PANs in federated settings with uneven data distributions. CoAtPENet, with its hybrid convolutional and attention-based architecture, emerges as the top-performing model, achieving an AUROC close to that of centralized learning (0.8230 vs. 0.8393). This result underscores CoAtPENet's ability to maintain high classification accuracy, even in a federated, imbalanced setup, and reflects the robustness in handling data variability common in medical image classification tasks.

Table 6.2 presents the results of different models on the multi-label classification task for the CheXpert dataset using the FedAvg algorithm, with evaluations across centralized, balanced, and imbalanced data distributions. In Table 6.2 (a) and Figure 6.10, we compare models with and without pre-training, while in table 6.2 (b), we incorporate PANs to analyze their impact on performance. Pre-training consistently improves the performance of all models, as evident from the AUROC scores (Fig. 6.10). For example, DenseNet121's AUROC improves from 0.8038 to 0.8132 in centralized settings and from 0.7612 to 0.7884 in imbalanced distributions. Similarly, CoAtNet benefits significantly from pre-training, achieving 0.8393 in centralized setups, the highest among all models, compared to 0.7873 without pre-training. MobileViT and CvT also show robust performance across all data distributions, with CvT achieving the highest scores (0.8453 in centralized, 0.8362 in balanced, and 0.8255 in imbalanced setups) among all models when pre-trained. Attention-based models like ConViT and CvT benefit more from pre-training compared to traditional CNNs, reflecting their ability to utilize global dependencies better, especially in imbalanced datasets.

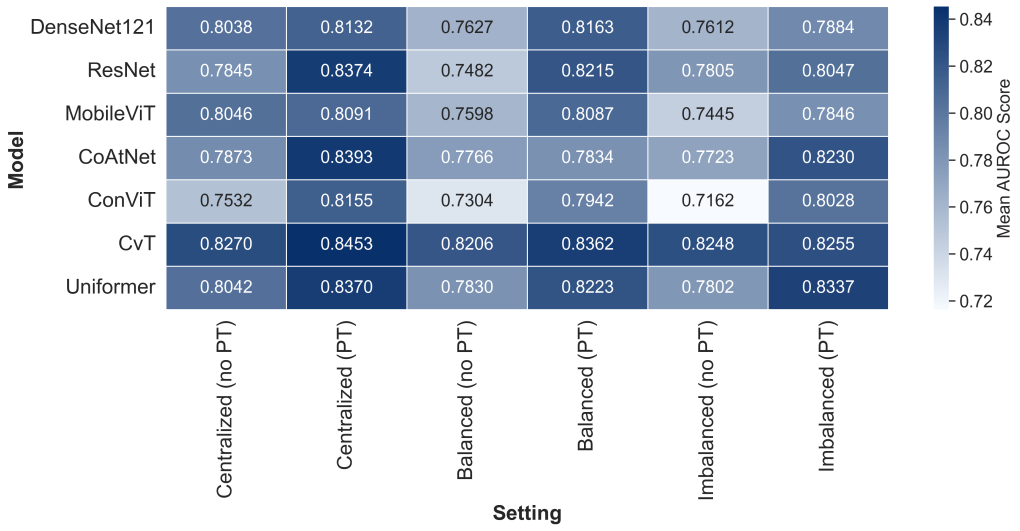


Figure 6.10: Mean AUROC Scores by Model and Setting of different DL models with and without pre-training on CheXpert.

Table 6.2: CheXpert dataset results on FedAvg algorithm.

Models	Pre-train	Mean AUROC scores		
		centralized	balanced	imbalanced
DenseNet121		0.8038	0.7627	0.7612
DenseNet121	✓	0.8132	0.8163	0.7884
ResNet		0.7845	0.7482	0.7805
ResNet	✓	0.8374	0.8215	0.8047
MobileViT		0.8046	0.7598	0.7445
MobileViT	✓	0.8091	0.8087	0.7846
CoAtNet		0.7873	0.7766	0.7723
CoAtNet	✓	0.8393	0.7834	0.8230
ConViT		0.7532	0.7304	0.7162
ConViT	✓	0.8155	0.7942	0.8028
CvT		0.8270	0.8206	0.8248
CvT	✓	0.8453	0.8362	0.8255
Uniformer		0.8042	0.7830	0.7802
Uniformer	✓	0.8370	0.8223	0.8337

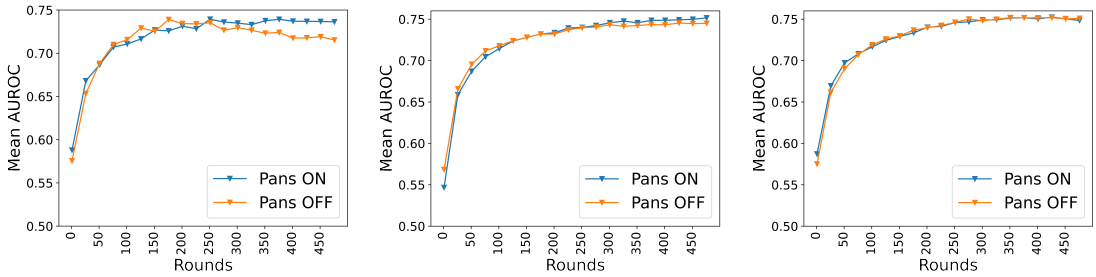
(a) Centralized and federated results of different models without PANs.

Models	Mean AUROC scores	
	balanced	imbalanced
Densenet121	0.7859	0.7723
MobileViT	0.7569	0.7783
CoAtPENet	0.7910	0.7947

(b) Federated results of different models with PANs.

Table 6.2 (a) and (b) present a comparative analysis of various DL models in the CheXpert dataset using the FedAvg algorithm under balanced and imbalanced conditions. Convolution-based models in 6.2 (a) reveal that pre-training models, such as ResNet and DenseNet121, perform better in balanced and imbalanced setups, achieving improved AUROC scores compared to their not pre-trained counterparts. DenseNet121, for instance, shows an increase in AUROC from 0.7612 to 0.7884 in the imbalanced setting with pre-training. Hybrid models such as ConViT, CvT, and Uniformer models exhibit varying degrees of effectiveness under both balanced and imbalanced setups, highlighting different strengths. Among these, Uniformer achieves the highest AUROC scores across both balanced and imbalanced settings, showing steady performance without the fluctuations observed in other architectures. This suggests that Uniformer's unique design, which blends convolutional and transformer elements, allows it to handle non-uniform data distributions effectively. Conversely, ConViT and CvT, while benefiting from pre-training, perform less consistently, par-

ticularly in imbalanced scenarios. ConViT, for example, shows limited adaptability under imbalanced conditions, with lower AUROC scores compared to models such as CoAtNet. This analysis indicates that while Uniformer and CoAtNet are more resilient to data imbalance, hybrid models like ConViT and CvT may require further optimization or complementary mechanisms like PANs to achieve similar robustness in federated medical imaging tasks. Table 6.2 (b) demonstrates the performance gains caused by integrating PANs, especially under imbalanced conditions, where CoAtPENet reaches a leading AUROC of 0.7947. IT indicates the effectiveness of PANs in enhancing model robustness, particularly for hybrid models like CoAtPENet, which is well suited to handle medical image classification challenges. The data collectively suggest that a combination of pre-training and PANs integration significantly boosts model performance under varying data distributions, making it a promising approach in FL for medical applications.



(a) DenseNet121 (b) MobileViT (c) CoAtNet

Figure 6.11: Imbalanced FedAvg results of DenseNet121, CoAtNet, and MobileViT on MIMIC-CXR dataset.

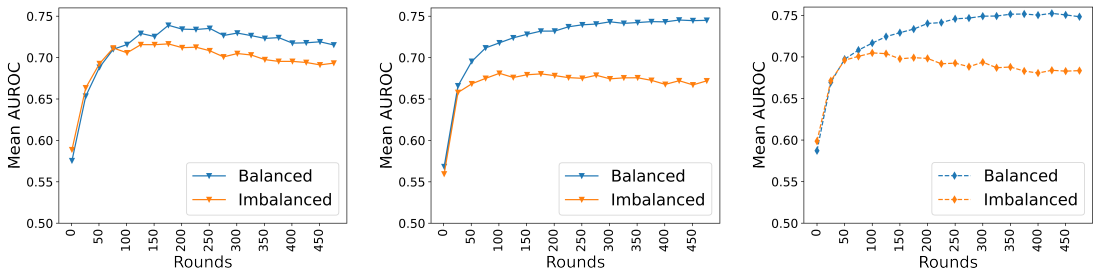
Figure 6.11 illustrates the performance of DenseNet121, MobileViT, and CoAtNet models using the FedAvg algorithm in the MIMIC-CXR dataset under FL conditions of imbalance. The results indicate that DenseNet121 with PANs shows the most substantial performance improvement, outperforming MobileViT and CoAtPENet in this configuration. It suggests that CNN-based architectures like DenseNet121 may benefit more from position-aware enhancements compared to attention-based models, such as MobileViT and CoAtPENet.

Figure 6.12 provides a comparative analysis of FedAvg performance in DenseNet121, CoAtPENet, and MobileViT models with PANs in different data distributions in the MIMIC-CXR dataset. It reveals that balanced data distributions consistently yield higher mean AUROC scores across all models, while the impact of PANs is particularly noticeable in the imbalanced setting, where they help reduce performance drop. CoAtPENet, in particular, demonstrates enhanced stability and accuracy, indicating that hybrid models benefit the most from such enhancements. These

findings underscore the effectiveness of PANs in FL, particularly for models such as CoAtPENet that take advantage of both spatial and positional encoding, thus improving performance in challenging and imbalanced scenarios.

Table 6.3 provides an in-depth comparison of various DL models in centralized and FL settings in both balanced and imbalanced data scenarios. The centralized setting in Table 6.3(a) demonstrates the superior performance of hybrid models such as CoAtNet and Uniformer, with high AUROC scores. CoAtNet achieves strong results because of its combination of convolutional and attention mechanisms, which allows for both spatial and global context modeling. Convolution-based models, such as ResNet and DenseNet, perform well in centralized setups, but show limitations when transitioning to federated settings, particularly under imbalanced conditions. This performance drop can be attributed to the reliance of convolutional models on spatial hierarchies, which can be disrupted when data distributions are uneven between FL clients.

When PANs are applied, as shown in Table 6.3(b), a noticeable improvement occurs in most models, especially in imbalanced configurations. The noticeable improvement can be seen in DenseNet, indicating that PANs help mitigate the effects of imbalanced data distribution by enhancing spatial context in FL settings. The attention-based MobileViT, shows only marginal drop with PANs, suggesting that PANs provide limited added benefit for models with built-in attention mechanisms. CoAtPENet achieves the little performance boosts, demonstrating resilience to imbalanced data with minimal accuracy degradation, and showing the highest AUROC in FL among hybrid models. This analysis underscores the potential of PANs in enhancing FL models, particularly in hybrid architectures that integrate convolution and attention, enabling them to perform more consistently in decentralized and imbalanced medical imaging applications.



(a) DenseNet121 (b) MobileViT (c) CoAtPENet

Figure 6.12: FedAvg results of DenseNet121, CoAtPENet, and MobileViT with PANs under different data distributions on MIMIC-CXR dataset.

To address the challenges posed by non-IID and imbalanced data in FL scen-

Table 6.3: MIMIC-CXR dataset results on FedAvg algorithm.

Models	Pre-train	Mean AUROC scores		
		centralized	balanced	imbalanced
DenseNet121		0.7422	0.7480	0.7176
DenseNet121	✓	0.7675	0.7527	0.7474
ResNet		0.7606	0.7504	0.7521
ResNet	✓	0.7716	0.7631	0.7663
MobileViT		0.7444	0.7546	0.7274
MobileViT	✓	0.7601	0.7490	0.7587
CoAtNet		0.7429	0.7390	0.7166
CoAtNet	✓	0.7699	0.7394	0.7639
ConViT		0.6417	0.6241	0.6415
ConViT	✓	0.6810	0.6802	0.6896
CvT		0.7712	0.7708	0.7697
CvT	✓	0.7725	0.7724	0.7707
Uniformer		0.6263	0.7025	0.7301
Uniformer	✓	0.7731	0.7809	0.7768

(a) Centralized and federated results of different DL models without PANs.

Models	Mean AUROC scores	
	balanced	imbalanced
Densenet121	0.7412	0.7210
MobileViT	0.7532	0.6872
CoAtPENet	0.7542	0.7179

(b) Federated results of different models with PANs.

arios, we explored two strategies aimed at narrowing the performance gap between centralized and federated settings: pre-training and the integration of PANs. Our results indicate that applying PANs and pre-training both yield similar performance improvements across the three datasets considered. However, PANs stand out by delivering comparable gains without certain drawbacks commonly associated with pre-training. Pre-trained DL models are typically trained on large, heterogeneous datasets, often spanning multiple domains. When these pre-trained models are applied to target datasets that differ substantially from their training domains, the transferred knowledge may be suboptimal. Additionally, pre-trained models can impose constraints on architectural flexibility. Adjusting the structure to meet specific target task requirements can become a complex endeavor, particularly if major architectural modifications are needed. In contrast, PANs improve the ability of a model to cope with non-IID and imbalanced data distributions without altering its underlying complexity. By introducing position-aware elements to the model

layers, PANs improve robustness and adaptability while preserving architectural flexibility. Consequently, on the basis of these findings, our subsequent experiments will focus on the use of PANs as the primary method for handling heterogeneous and imbalanced data conditions.

6.2 Performance Evaluation for Different FL Algorithms

6.2.1 CovidX Dataset

Table 6.4 and Figure 6.13 collectively highlight the impact of different FL algorithms (FedAvg and FedProx) and the use of PANs on model performance for the CovidX dataset in non-IID settings. From Table 6.4, PANs consistently improve both accuracy and F1 scores across all models and algorithms, underscoring their role in mitigating neuron misalignment in FL. CoAtPENet achieves the highest accuracy and F1 score with FedAvg, outperforming other configurations. MobileViT also performs well with FedAvg and PANs, achieving 89.75% accuracy and a slightly lower F1 score of 0.82. In contrast, DenseNet121 shows modest improvements, achieving 81% accuracy with FedProx and PANs, but lags behind CoAtPENet and MobileViT in overall performance. Without PANs, accuracy and F1 scores decline across all setups, with the drop being more pronounced in DenseNet121, further emphasizing the importance of PANs for enhancing generalization in non-IID FL scenarios.

Figure 6.13 provides a dynamic perspective on trends in 100 communication rounds, revealing that FedAvg with PAN leads to the highest and most stable accuracy, particularly for CoAtPENet, as it converges around the accuracy 90% after 40 rounds. FedProx with PANs also shows significant stability, though its peak accuracy is slightly lower than FedAvg with PANs. In contrast, configurations

Table 6.4: Results of different FL algorithms with different DL models under non-IID setting for CovidX dataset.

Models	FL algorithms	PANs on		PANs off	
		Acc. (%)	F1 score	Acc. (%)	F1 score
DenseNet121	FedAvg	75.70	0.80	69.50	0.60
	FedProx	81	0.76	76	0.75
MobileViT	FedAvg	89.75	0.82	82	0.81
	FedProx	79.5	0.75	75	0.71
CoAtPENet	FedAvg	91.10	0.84	85	0.83
	FedProx	86.50	0.84	83	0.81

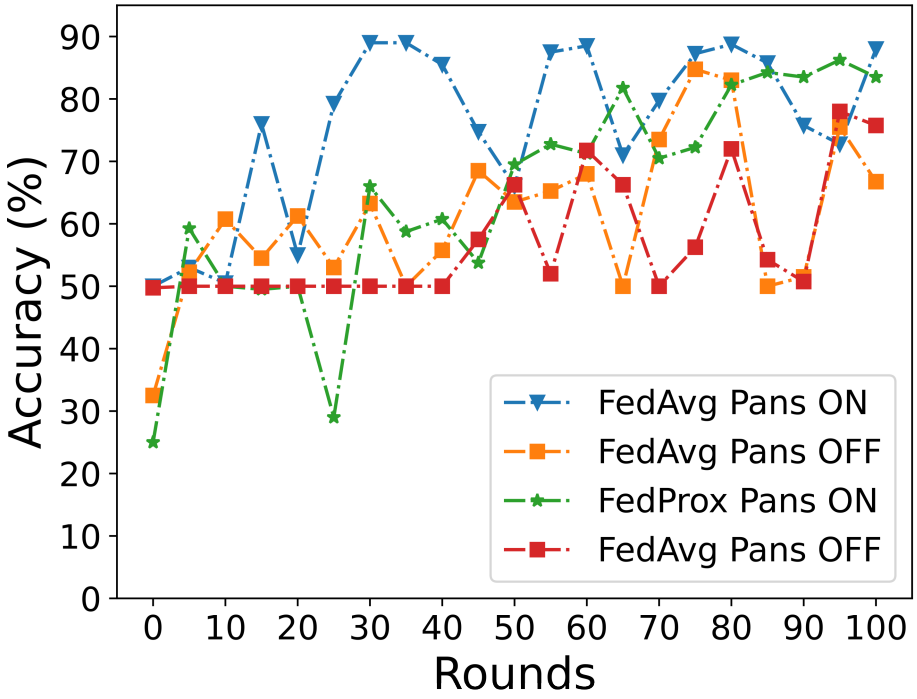


Figure 6.13: CoAtPENet with various FL algorithms (CovidX).

without PANs exhibit more variability and lower convergence, with FedAvg and FedProx plateauing at much lower accuracy levels. A clear pattern emerges when PANs consistently enhance learning stability and final accuracy for both algorithms, while FedAvg slightly outperforms FedProx in terms of peak accuracy when PANs are enabled. These results highlight the synergistic effect of using PANs with FL algorithms, particularly FedAvg, to achieve superior performance in non-IID FL setups.

6.2.2 CheXpert and MIMIC-CXR Datasets

The CheXpert dataset, evaluated using different FL algorithms in imbalanced settings, highlights the impact of PANs and FL approaches on the AUROC performance. Table 6.5 and Figure 6.14 show that CoAtPENet achieves the highest performance, reaching a mean AUROC of 0.8118 under the FedProx algorithm, outperforming other models and configurations. DenseNet121 and MobileViT also see improvements with PANs, with FedProx enhancing their scores to 0.8061 and 0.7629, respectively. However, FedAvg consistently delivers lower AUROC values compared to FedProx, indicating that FedProx's stabilization mechanisms are more effective in addressing client variability and data heterogeneity. Figure 6.15 further emphasizes this, as CoAtPENet with FedProx (green line) maintains the highest AUROC and the most stable convergence across 500 rounds, while configurations without PANs show

Table 6.5: Results of different FL algorithms with different DL models under imbalanced setting for CheXpert dataset.

Models	FL algorithms	Mean AUROC scores	
		PAN on	PAN off
DenseNet121	FedAvg	0.7939	0.7446
	FedProx	0.8061	0.7744
MobileViT	FedAvg	0.7783	0.7612
	FedProx	0.7629	0.7411
CoAtPENet	FedAvg	0.7946	0.7445
	FedProx	0.8118	0.7714

Table 6.6: Results of different FL algorithms with different DL models under imbalanced setting for MIMIC-CXR dataset.

Models	FL algorithms	Mean AUROC scores	
		PAN on	PAN off
DenseNet121	FedAvg	0.7412	0.7390
	FedProx	0.7531	0.7451
MobileViT	FedAvg	0.7532	0.7481
	FedProx	0.7534	0.7517
CoAtPENet	FedAvg	0.7547	0.7542
	FedProx	0.7556	0.7540

noticeable oscillations and lower final AUROC values. The results clearly demonstrate that both PANs and FedProx play crucial roles in optimizing FL for multi-label classification tasks in medical datasets.

However, the results of the MIMIC-CXR dataset, presented in Table 6.6 and Figures 6.16 and 6.17, reinforce the patterns observed in CheXpert. CoAtPENet again achieves the highest mean AUROC scores, with 0.7556 under FedProx, closely followed by FedAvg (0.7547). DenseNet121 and MobileViT exhibit similar trends, with FedProx improving their AUROC to 0.7531 and 0.7532, respectively, when PANs are used. Figures 6.16 show that all configurations converge by 500 rounds, but FedProx with PANs provides more stability and slightly higher AUROC compared to other combinations. In particular, the inset zoomed plot highlights that PANs reduce variability and improve final performance for all models, particularly in the presence of data imbalance. Compared to others, FedAvg without PAN consistently underperforms, plateauing at lower AUROC values.

These findings indicate that both FL algorithms can achieve strong performance when applying PANs to MobileViT, DenseNet121, and CoAtNet DL models on the CheXpert and MIMIC-CXR datasets.

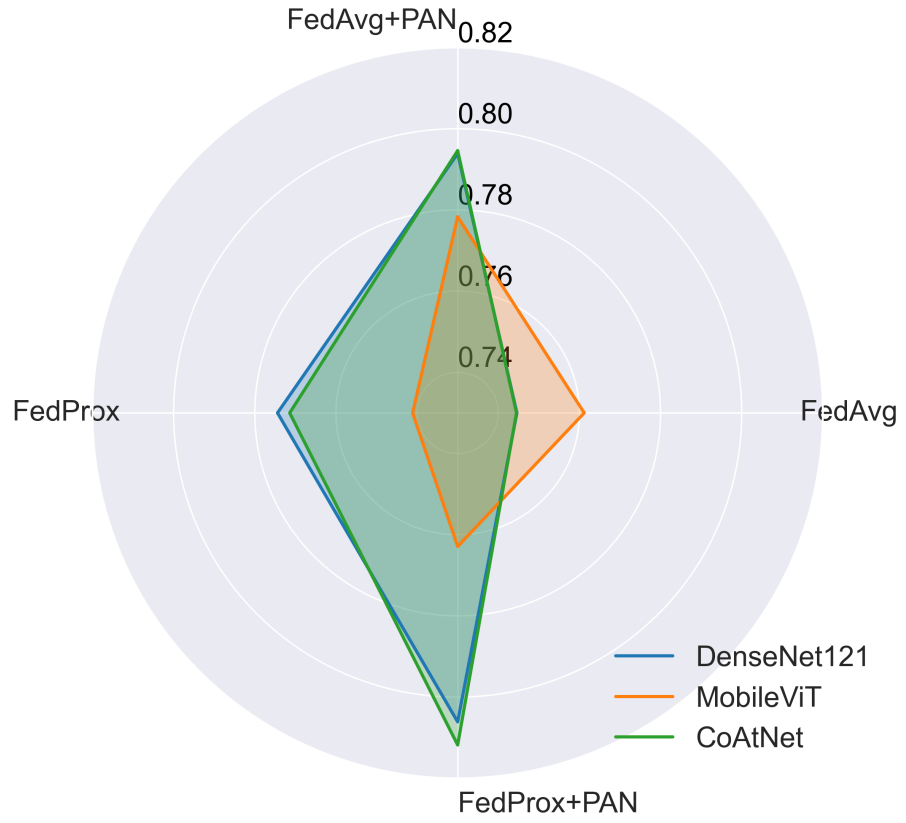


Figure 6.14: CheXpert AUROC Score Comparison by Algorithm and PAN Setting.

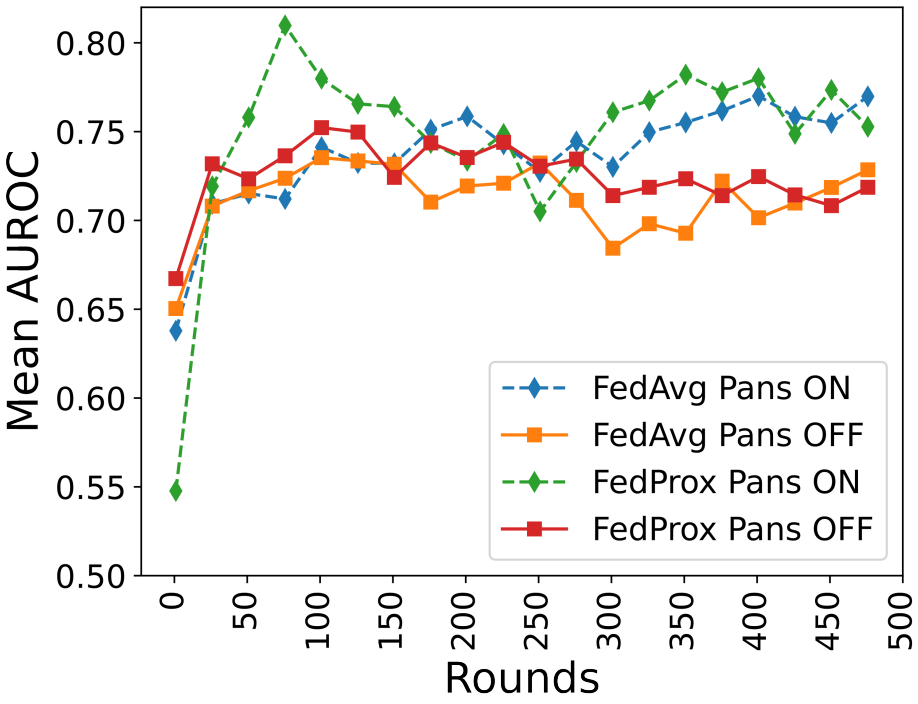


Figure 6.15: CoAtPENet with various FL algorithms (CheXpert).

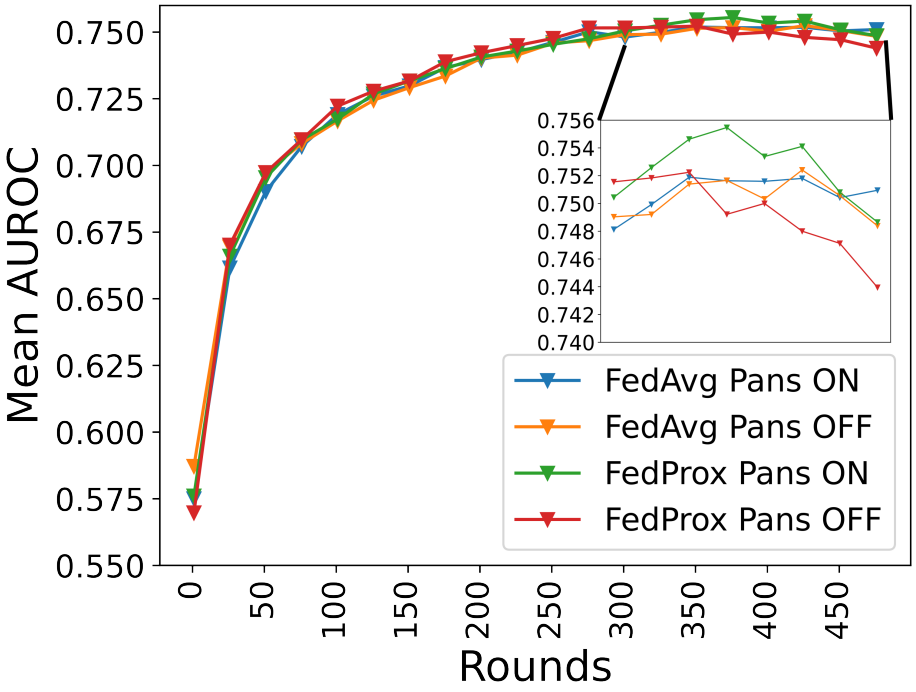


Figure 6.16: CoAtPENet with various FL algorithms (MIMIC-CXR).

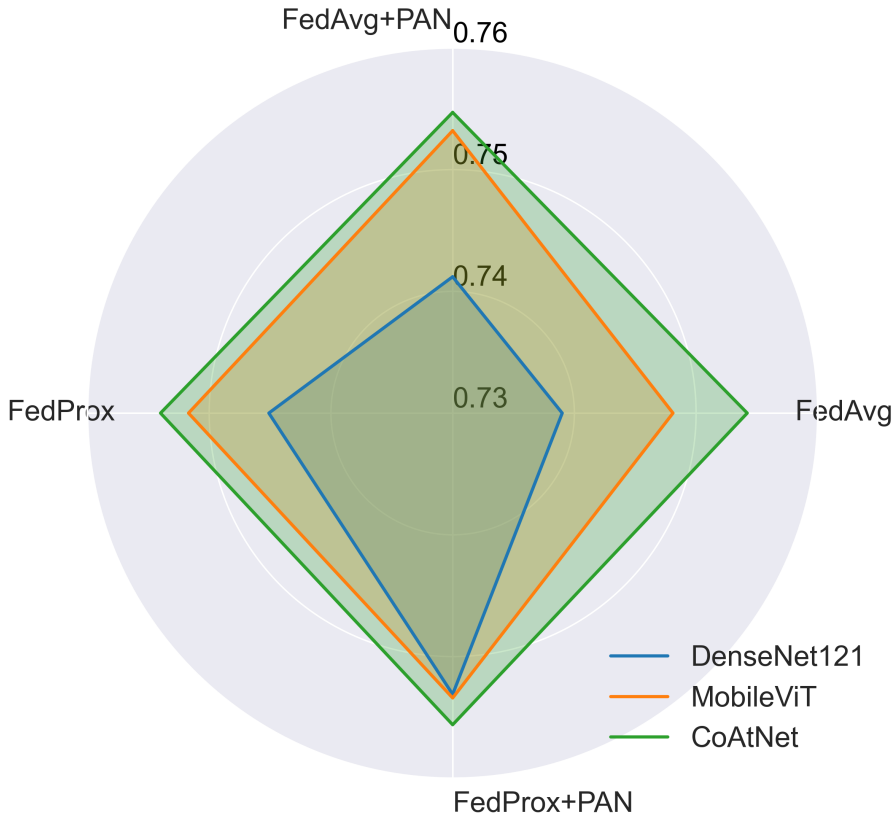


Figure 6.17: MIMIC-CXR AUROC Score Comparison by Algorithm and PAN Setting.

6.3 Effect on Different Number of Clients

6.3.1 CovidX Dataset

This section explores how varying the number of clients affects performance under the FedAvg algorithm on a non-IID CovidX dataset. As the number of clients in our FL system increases from 2 to 20, we observe a significant decline in model accuracy due to the fragmentation of the entire data created by the LDA distribution method introduced in section 4.2.1. In this approach, each client receives a unique mix of classes based on probability vectors sampled from a Dirichlet distribution with parameter α, which determines how non-uniformly the classes are distributed. When we scale to more clients (10 or 20), the same dataset is split into smaller and more specialized subsets, often creating imbalances in class representation. It produces non-IID conditions where some clients may have an abundance of certain classes while completely lacking others. This fragmented distribution makes it increasingly difficult for the global model to reconcile the diverse learning patterns of each client, explaining why performance degrades as the number of participating clients increases.

Figure 6.18 explores the impact of varying the number of clients (2, 5, 10, and 20) on the accuracy of CoatPENet using the FedAvg algorithm using the CovidX dataset. In this experiment, the number of images per client is as follows:

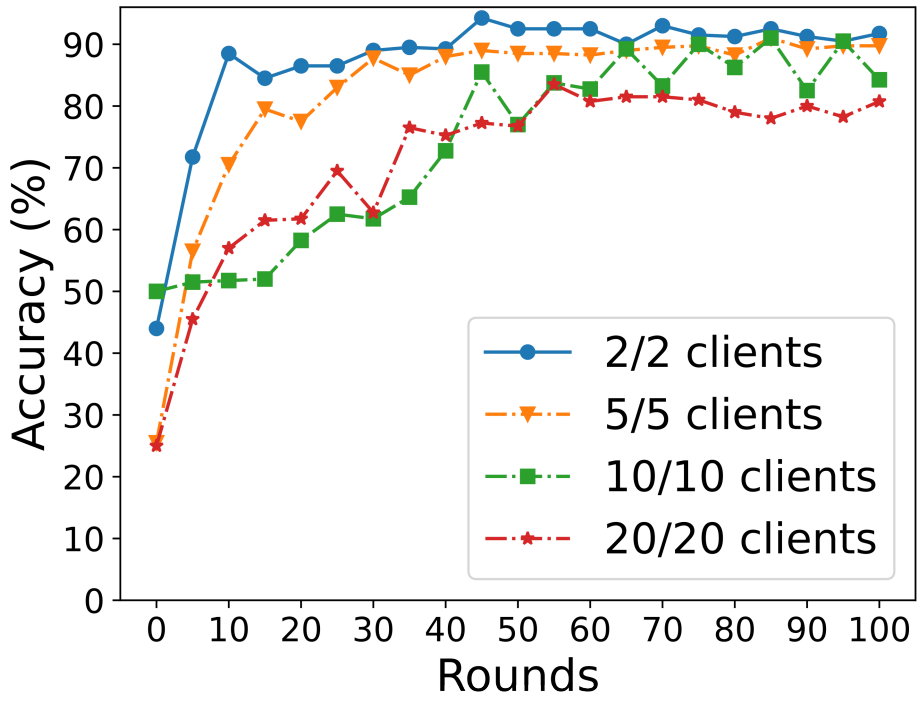


Figure 6.18: CoatPENet with different number of clients under FedAvg (CovidX).

- 2/2 clients (6987 images per client)
- 5/5 clients (2795 images per client)
- 10/10 clients (1397 images per client)
- 20/20 clients (698 images per client)

The results show a clear relationship between the number of clients, the accuracy, and the convergence speed. With fewer clients (2/2 and 5/5), the model converges rapidly, achieving 90% accuracy within 30 rounds for 2/2 clients and 85% for 5/5 clients. This can be attributed to the reduced heterogeneity and more data available per client, resulting in more stable and efficient updates.

As the number of clients increases to 10/10 and 20/20, convergence slows and the final accuracy drops to 80% and 75%, respectively. The larger number of clients introduces greater heterogeneity and smaller local datasets per client, which negatively affect model aggregation and learning stability. The fluctuations in the 20/20 client setup are particularly pronounced, reflecting the challenges of balancing updates across diverse and data-constrained clients in a federated learning setting. These findings highlight the trade-offs between scalability and performance, where fewer clients enable faster convergence and higher accuracy but limit scalability, while more clients increase FL scalability at the cost of reduced accuracy and stability.

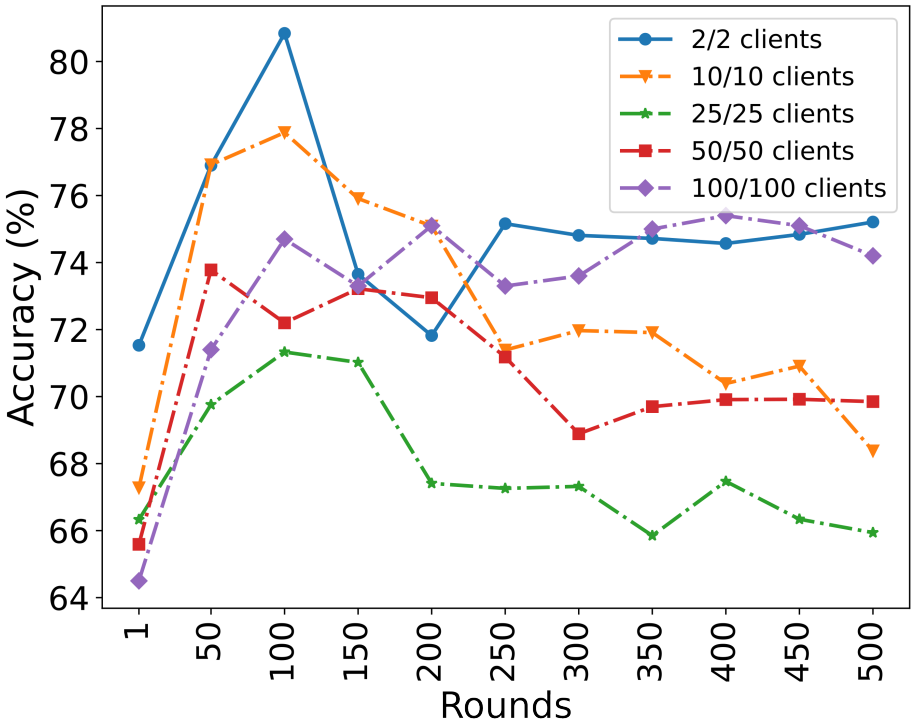


Figure 6.19: CoatPENet with different number of clients under FedAvg (CheXpert).

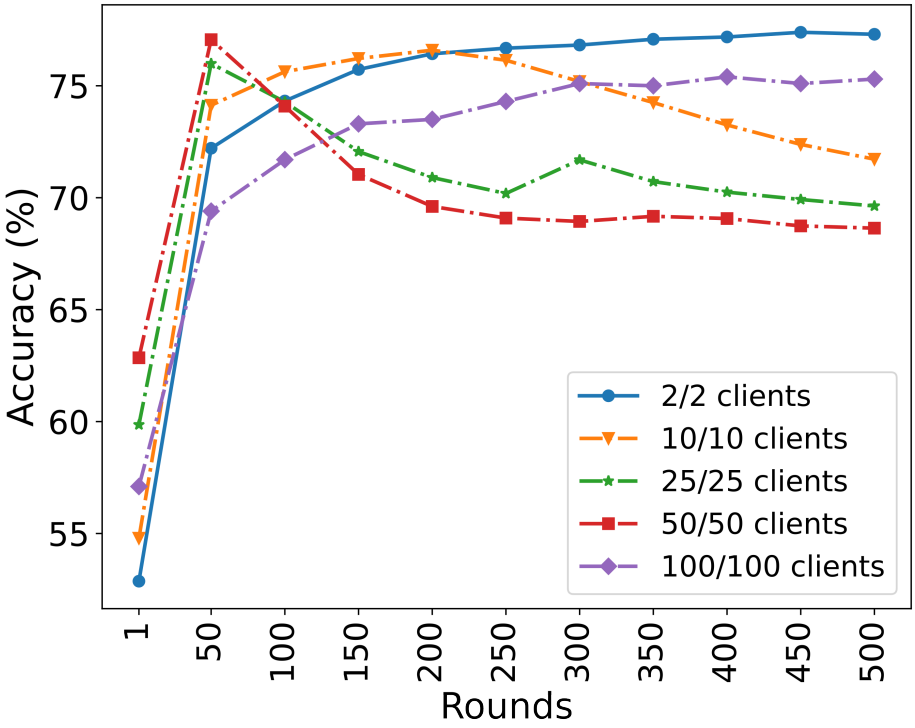


Figure 6.20: CoatPENet with different number of clients under FedAvg (MIMIC-CXR).

6.3.2 CheXpert and MIMIC-CXR Datasets

In this section, we examine how varying the number of clients affects performance on multi-label datasets. Figure 6.19 displays the accuracy trends of the CoAtPENet model under different client configurations for the CheXpert dataset. When using only 2 clients, the system attains the highest accuracy, reaching nearly 80% early on and later stabilizing around 74%. In contrast, as the number of clients grows, accuracy generally diminishes. With 100 clients, the model performance is the lowest, hovering between 66-68% after some initial fluctuations. The 10- and 25-client setups yield better results than larger client counts, reaching a peak near 76% before gradually decreasing. However, even with this decline, they maintain a higher accuracy than in the 50- and 100-client scenarios. These observations suggest that the use of fewer clients facilitates a more effective aggregation of model updates, resulting in more stable and accurate results for the CheXpert dataset.

Figure 6.20 presents similar results for the MIMIC-CXR dataset, although the changes are more gradual. As before, the 2-client configuration achieves the best accuracy, stabilizing around 75%. Alternative configurations involving 10, 25, and 50 client clusters exhibit results within the range of 70-73%. Although the 50-client setup reaches approximately 72% at its peak and then experiences a slight decrease, the gap between different client counts is narrower than in CheXpert. This indicates

that MIMIC-CXR is less sensitive to the number of clients, possibly due to differences in data complexity or distribution characteristics. Unlike CheXpert, the decline in accuracy with increasing rounds is not as pronounced, and performance levels off earlier.

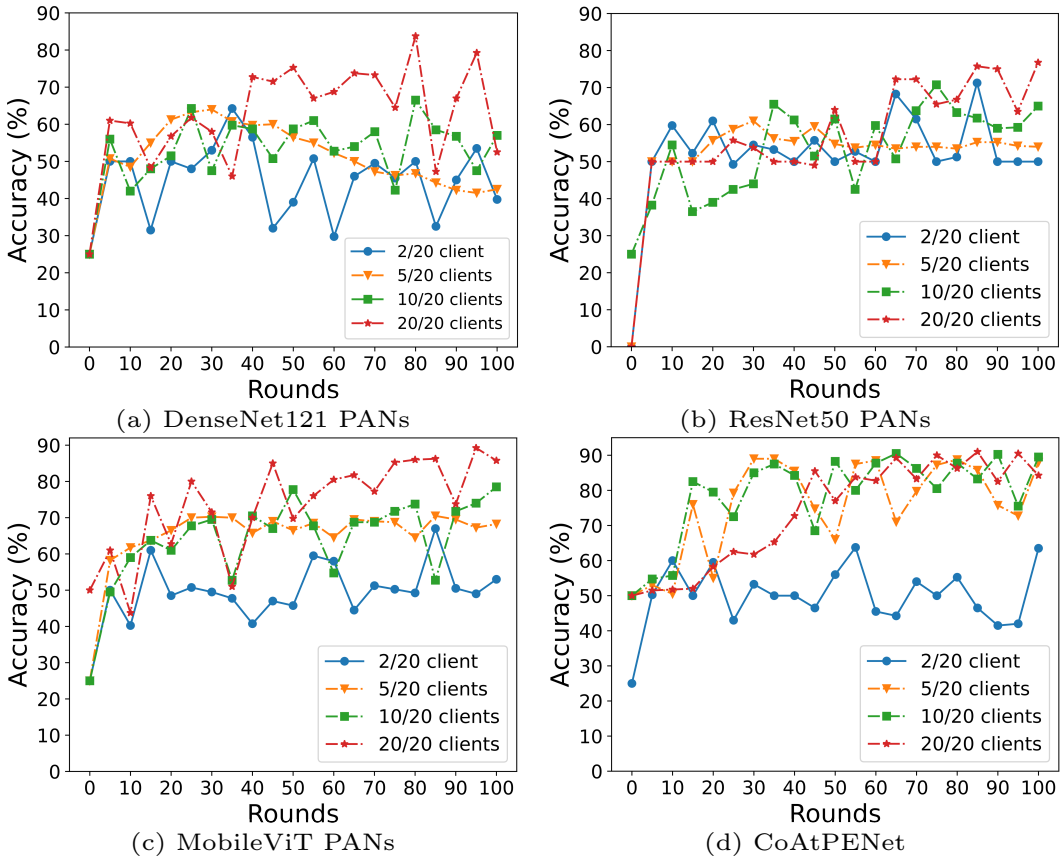
Across both datasets, increasing the number of clients tends to reduce overall performance. This trend may be attributed to the complexities inherent in multi-label tasks, where model divergence or noisier updates are more likely. Smaller client groups benefit from more stable updates and quicker convergence, enhancing performance. However, the magnitude of these effects varies by dataset: CheXpert exhibits a more significant performance gap between configurations than MIMIC-CXR, illustrating how dataset-specific factors can influence the optimal balance of client numbers to maximize FL outcomes.

6.4 Effect on Changing the Percentage of Participants

6.4.1 CovidX Dataset

In these experiments, we investigate how changing the percentage of participating clients affects performance on the CovidX dataset. In figure 6.21, we show the results for several DL models integrated with PANs under the FedAvg algorithm. In Figure 6.21(a), DenseNet121 with PANs exhibits substantial accuracy fluctuations, varying between 30% and 80%, indicating a high sensitivity to the proportion of data utilized. In contrast, ResNet50 performance (Fig. 6.21 (b)) shows a notable improvement as the participation ratio increases from 10% to 100%, with accuracy advancing up to 70.5%. With full participation, the model stabilizes after approximately 50 rounds, reaching a maximum accuracy of 76. 5%. However, when fewer participants are involved (2 or 5 clients), the model performance remains near 50%, interrupted only by occasional surges.

MobileViT with PANs (Fig. 6.21(c)) exhibits more consistent learning curves compared to CNN-based models. Although participation in 100% produces the highest accuracy (90%), reducing the ratio to 50% cuts the accuracy by approximately 15% (falling to 79%). With fewer participants, performance varies widely and never exceeds 70%. However, CoAtPENet (Fig. 6.21(d)) demonstrates robust performance, maintaining accuracies above 90% even when limited to smaller participant subsets (5, 10, or 20 clients). This outcome highlights the strength of the hybrid architecture in adapting effectively to different participation levels in federated learning.



(a) DenseNet121 PANs (b) ResNet50 PANs

(c) MobileViT PANs (d) CoAtPENet

Figure 6.21: FL results of 4 models with PANs when changing the percentage of participants on CovidX dataset.

6.4.2 CheXpert and MIMIC-CXR Datasets

Figure 6.22 illustrates the effect of changing the percentage of participating clients on the performance of MobileViT with PANs and CoAtPENet, for the imbalanced CheXpert dataset. The results are reported in terms of mean AUROC over 500 rounds, highlighting the interaction between client participation rates and FL outcomes.

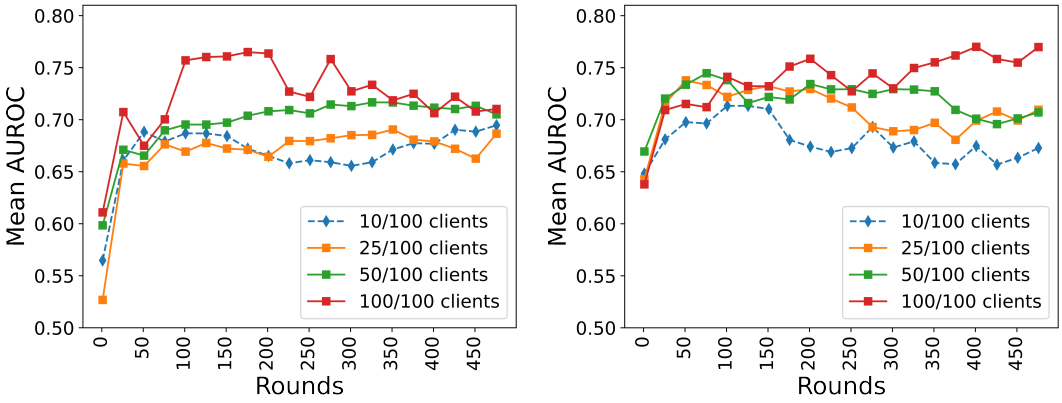
In Figure 6.22 (a), for the MobileViT model, as the number of participating clients increases, the mean AUROC decreases and exhibits more variability. When only 10 clients participate, the model achieves the highest AUROC with relatively stable convergence. However, with 25, 50, and 100 participating clients, the AUROC progressively drops, reaching 0.65 for 100 clients. This degradation is due to increased client heterogeneity and smaller local datasets for each participant, which introduce greater challenges for model aggregation in FL. For CoAtPENet in Figure 6.22 (b), a similar trend is observed, but CoAtPENet shows better resilience to increased client participation compared to MobileViT. With 10/100 clients, CoAtPENet reaches a peak AUROC of 0.78, maintaining higher accuracy than MobileViT in this setup. As participation increases to 25/100 and 50/100, CoAtPENet retains relatively

stable performance, achieving 0.75 and 0.72, respectively. Even with 100/100 clients, CoAtPENet maintains an AUROC close to 0.70, showing its robustness in handling higher client variability and smaller local datasets.

For MobileViT (Figure 6.23 (a), the performance remains consistent with all the client participation rates, with mean AUROC values converging around 0.75 by 450 rounds. Although there are slight variations during training, the differences between configurations (e.g., 10/100 and 100/100) are negligible, indicating that MobileViT is less sensitive to participation rates in this dataset. The inset plot highlights minor fluctuations, with the 10/100 configuration showing marginally higher stability and peak AUROC (0.751). CoAtPENet in Figure 6.23 (b), shows a similar trend, with all configurations converging near 0.754 at the end of training. CoAtPENet demonstrates a slight advantage over MobileViT in terms of final AUROC and stability. The inset plot reveals that 10/100 clients achieve the highest mean AUROC during intermediate rounds (0.754), while 100/100 clients exhibit slightly more variability, although the final performance difference is minimal.

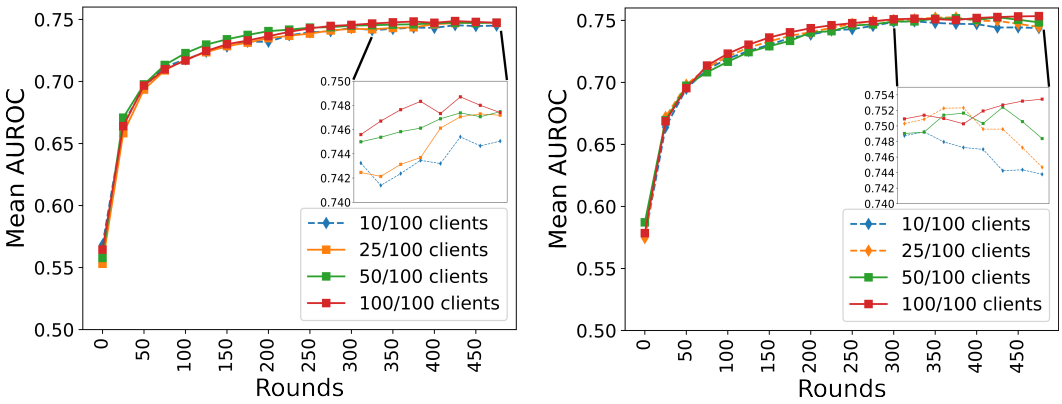
These results reveal a trade-off between participation rates and model performance in FL. While lower participation ensures higher local data availability and stability, higher participation introduces diversity at the cost of performance. CoAtPENet proves to be more robust than MobileViT under increased participation, making it a better choice for scenarios that require greater client involvement in FL setups. For the CheXpert dataset, CoAtPENet consistently outperforms MobileViT under varying client participation rates, demonstrating greater resilience to the challenges introduced by client heterogeneity and imbalanced data. MobileViT, while effective in scenarios with fewer participating clients, exhibits a more pronounced drop in performance as participation increases.

For the MIMIC-CXR dataset, both MobileViT and CoAtPENet exhibit high robustness to changes in client participation, with PANs effectively mitigating the impact of increased client heterogeneity. CoAtPENet achieves slightly better mean AUROC and stability than MobileViT, making it the more reliable choice for this multi-label classification task, especially in highly distributed settings. The minimal performance drop observed with increased participation in MIMIC-CXR further underscores the ability of PANs to handle imbalanced data and the federated learning setup effectively. These findings confirm CoAtPENet's superior adaptability across datasets and participation rates, positioning it as the preferred architecture for complex, distributed FL scenarios.



(a) MobileViT (b) CoAtPENet

Figure 6.22: MobileViT with PANs and CoAtPENet enabled results when changing the percentage of participants on imbalanced CheXpert dataset.



(a) MobileViT (b) CoAtPENet

Figure 6.23: MobileViT with PANs and CoAtPENet enabled results when changing the percentage of participants on imbalanced MIMIC-CXR dataset.

6.5 Comparison with State-of-the-Art

This section compares our approach with state-of-the-art (SOTA) results. We evaluated different methods based on various factors, including the chosen FL algorithm, the model architecture, data distribution scenarios (IID vs. non-IID), and whether pre-training or PANs are employed. In the tables, the best performing result is emphasized in **bold**, while the second best is underlined.

Table 6.7 compares the performance of the proposed methods with the SOTA approaches on the CovidX dataset under various configurations. Among SOTA methods, ResNext achieves the best accuracy (92%) but does not incorporate non-IID settings or PANs, limiting its applicability in challenging FL scenarios. ResNet50, evaluated with and without pre-training in non-IID settings, achieves slightly lower results 90.86% and 91.49%, respectively, demonstrating its competitiveness, but highlighting the importance of pre-training for improved performance in federated setups.

Our proposed approach using CoAtPENet achieves comparable or better results with significant versatility across configurations. CoAtPENet matches ResNext performance (92%) under IID conditions without PANs, showcasing the strength of its hybrid architecture even in simpler setups. More importantly, CoAtPENet maintains strong results in non-IID settings, achieving 91% accuracy without pre-training and PANs. This demonstrates the robustness of the proposed method in handling client heterogeneity and limited data per client, areas where traditional SOTA models falter.

In general, Table 6.7 illustrates that while SOTA methods such as ResNext and ResNet50 deliver competitive results under specific conditions, the proposed CoAtNet-based solutions provide flexibility and consistent performance in a wider range of FL scenarios, particularly in non-IID and pre-training setups. This highlights the advantages of integrating PANs and hybrid model architectures to improve generalization and robustness in federated learning tasks.

Table 6.8 provides a detailed comparison of the proposed approach with the SOTA methods in the CheXpert dataset, focusing on the best mean AUROC as an evaluation metric in various FL configurations. Among the SOTA methods, ResNet34 achieves the highest mean AUROC of 0.8294 under the FedDF algorithm, demonstrating its strength in this specific setup. However, this approach lacks pre-training and does not explicitly account for data imbalance, which may limit its generalizability in more realistic FL scenarios. Similarly, ResNet34 achieves the second-best performance (0.8262) under FedAvg with pre-training, showcasing its adaptability when pre-trained.

Table 6.7: CovidX dataset results comparison with other works.

Method	FL Algorithm	Model	Non-iid	Pre-training	PAN	Results (Best Acc.%)
[63]	FedAvg	ResNext	x	x	x	**92**
[120]	FedAvg	ResNet50	✓	x	x	90.86
[120]	FedAvg	ResNet50	x	x	x	<u>91.49</u>
Ours	FedAvg	CoAtNet	x	✓	x	**92**
	FedAvg	CoAtNet	✓	✓	x	89
	FedAvg	CoAtNet	x	x	x	**92**
	FedAvg	CoAtNet	✓	x	x	85
	FedAvg	CoAtNet	x	x	✓	91
	FedAvg	CoAtNet	✓	x	✓	**92**

Note: The best result is **bold** and the second-best result is <u>underlined</u>.

ViT, another SOTA model, performs well under FedAvg with pre-training, achieving a mean AUROC of 0.814, but it still lags behind ResNet34. In comparison, our CoAtNet models deliver highly competitive results, particularly when augmented with PANs and using the FedProx algorithm. CoAtNet achieves a mean AUROC of 0.8118 under FedProx without pre-training and PANs, making it the most effective among the proposed configurations for handling imbalanced data in FL settings. Without PANs, CoAtNet maintains strong performance, achieving a mean AUROC of 0.8230 under FedAvg in setups with pre-training. Although ResNet34 slightly outperforms CoAtNet in select configurations, the proposed CoAtNet demonstrates remarkable adaptability and reliability in multiple configurations, benefiting significantly from innovations such as PANs and the FedProx algorithm. These findings confirm that CoAtNet is a highly competitive and versatile architecture for FL in multi-label classification tasks, making it a strong candidate for real-world applications where imbalanced data and non-IID distributions are common.

Table 6.9 compares the proposed methods with the SOTA approaches in the MIMIC-CXR dataset, highlighting the best mean AUROC as the primary metric under various FL configurations. Among the SOTA methods, ViT achieves the best performance, with a mean AUROC of 0.833 under the FedAvg algorithm, leveraging pre-training to handle imbalanced data effectively. DenseNet121, evaluated under the FedFBN algorithm, achieves a mean AUROC of 0.75, showing competitive results but falling short of ViT due to its limited adaptability to the complexities of federated setups.

The proposed CoAtPENet models demonstrate consistent and competitive results in various configurations on the MIMIC-CXR dataset. Without PANs and no

Table 6.8: CheXpert dataset results comparison with other works.

Method	FL Algorithm	Model	Imbalanced	Pre-training	PAN	Results (Best mean AUROC)
[11]	FedAvg	ResNet18	x	✓	x	0.79
[51]	FedFBN	DenseNet121	✓	✓	x	0.75
[105]	FedAvg	ViT	✓	✓	x	0.814
[31]	FedAD	ResNet34	x	NA	x	<u>0.8262</u>
[31]	FedMD	ResNet34	x	NA	x	0.7766
[31]	FedDF	ResNet34	x	NA	x	**0.8294**
Ours	FedAvg	CoAtNet	x	✓	x	0.7832
	FedAvg	CoAtNet	✓	✓	x	0.8230
	FedAvg	CoAtNet	x	x	x	0.7766
	FedAvg	CoAtNet	✓	x	x	0.7723
	FedAvg	CoAtNet	x	x	✓	0.7910
	FedAvg	CoAtNet	✓	x	✓	0.7947
	FedProx	CoAtNet	✓	x	✓	0.8118

Note: The best result is **bold** and the second-best result is <u>underlined</u>.

Table 6.9: MIMIC-CXR dataset results comparison with other works.

Methods	FL Algorithm	Model	Imbalanced	Pre-training	PAN	Results (Best mean AUROC)
[105]	FedAvg	ViT	✓	✓	x	**0.833**
[51]	FedFBN	DenseNet121	✓	✓	x	0.75
Ours	FedAvg	CoAtNet	x	✓	x	0.7394
	FedAvg	CoAtNet	✓	✓	x	<u>0.7639</u>
	FedAvg	CoAtNet	x	x	x	0.7390
	FedAvg	CoAtNet	✓	x	x	0.7166
	FedAvg	CoAtNet	x	x	✓	0.7079
	FedAvg	CoAtNet	✓	x	✓	0.7542
	FedProx	CoAtNet	✓	x	x	0.7540
	FedProx	CoAtNet	✓	x	✓	0.7556

Note: The best result is **bold** and the second-best result is <u>underlined</u>.

pre-training, CoAtNet achieves modest mean AUROCs of 0.709 and 0.7166 under the FedAvg algorithm, reflecting its baseline capability in FL tasks. Pre-training significantly enhances CoAtNet's performance, with mean AUROCs increasing to 0.7390 and 0.7639 under FedAvg, highlighting the importance of leveraging pre-trained weights for generalization across imbalanced datasets. The second highest performance comes from CoAtPENet with FedProx. Under this setup, comes from achieves a mean

AUROC of 0.7556, showcasing its ability to manage data heterogeneity and imbalance effectively while maintaining stable performance across clients.

Although ViT achieves the highest mean AUROC of 0.833 under the FedAvg algorithm, this result is heavily dependent on pre-training and does not incorporate PANs. CoAtPENet's performance with FedProx, while lower than ViT, demonstrates greater adaptability in realistic non-IID federated setups, making it a robust choice for FL applications. These results underscore the impact of architectural enhancements such as PANs and the stabilization provided by FedProx in optimizing performance for medical imaging datasets, particularly in environments with imbalanced data and diverse client distributions.

6.6 Computation and Communication Efficiency Analysis

In FL, computation and communication efficiency play a critical role in determining the feasibility of the model for deployment on edge devices. This section evaluates the computational complexity, model size, and communication costs of the proposed CoAtPENet model compared to baseline models. The goal is to provide information on resource requirements in practical FL deployments, particularly in resource-constrained environments.

6.6.1 Model Size and Parameter Complexity

In Table 6.10, we show the critical balance between model size, training time, and inference speed when selecting DL architectures for FL applications. The results reveal the trade-offs between computational efficiency and model performance, highlighting the unique strengths and limitations of different architectures when

Table 6.10: Model size, training time, and inference on CovidX dataset.

Models	# of parameters (M)	Total training time (s)	Inference (s)
DenseNet121	8	196408.4	0.0171 ± 0.0002
ResNet50	23.5	193596.4	0.0072 ± 0.0001
ViT	86	437596.61	0.0653 ± 0.0006
MobileViT	5.6	200196.80	0.0089 ± 0.0004
CoAtNet	16.9	199187.8	0.0113 ± 0.0004
ConViT	27.2	204627.0	0.0114 ± 0.0004
CvT	20	208372.2	0.0208 ± 0.0009
Uniformer	23.5	219577.8	0.0190 ± 0.0005

applied to the CovidX dataset. Initially, we explored the state-of-the-art ViT base model, recognized for its strong performance. As Dosovitskiy et al. [18] suggest, the ViT base model, with its substantial 86M parameter size, is ideally suited for large datasets. Although our study included large datasets such as CheXpert and MIMIC-CXR, the CovidX dataset was medium-sized, comprising 20,000 images. Furthermore, in our FL setup, the CovidX data was distributed across 20 clients, resulting in an average of only 2,000 images per client. Consequently, training the ViT base model led to significant fluctuations and prolonged training times. This prompted us to shift our focus to more lightweight models, such as MobileViT. Among the models, MobileViT emerges as the most computationally efficient, with the smallest parameter size (5.6M) and a relatively fast inference time. Despite its lightweight design, MobileViT maintains competitive performance, making it an excellent choice for resource-constrained FL environments. DenseNet121 has a modest parameter size (8M) and a slightly longer inference time, but its total training time is shorter compared to larger models, reflecting its efficiency for moderate-scale deployments.

In contrast, hybrid models such as ConViT and CvT exhibit longer training times and slower inference speeds. This highlights their computational intensity and suitability for scenarios where accuracy takes precedence over speed. Uniformer demonstrates a balance between complexity and inference time, but remains computationally expensive during training, reflecting its challenges in resource-constrained FL setups.

CoAtNet, with 16.9M parameters, offers a compelling balance between computational efficiency and performance. Its total training time is moderate compared to larger hybrid models, while its inference time is reasonable, indicating its architecture's ability to integrate efficiency with high performance. ResNet50 having the same parameter size as Uniformer achieves the fastest inference time, demonstrating the efficiency of traditional CNN-based architectures, though it sacrifices adaptability in non-IID federated settings.

When comparing DenseNet vs CoAtNet, then it is evident that the accuracy over model size matters (77% vs. 92%). In the case of MobileViT versus CoAtNet, we advocate that an increase in accuracy 2% can be highly significant considering medical image analysis. Even a small improvement can lead to earlier or more accurate diagnoses, potentially resulting in better patient outcomes. Furthermore, our proposed CoAtPENet model can potentially be optimized in the future. Techniques such as model pruning, quantization, or knowledge distillation can be used to reduce the size of the model without significantly sacrificing accuracy. So, the current size

of our model is not necessarily a permanent limitation.

Table 6.11: Trainable parameters and FLOPs per image comparison.

Model	**FLOPs**	**Params (M)**
DenseNet	2.9G	8
ResNet	4.1G	23.5
MobileViT	1.7G	5.6
CoAtPENet	3.4G	16.9

Table 6.12: GPU and RAM memory consumption of different DL models.

Model	**Training**		**Inference**	
	RAM	GPU	RAM	GPU
DenseNet	2.58Gb	6.25Gb	2Gb	1.75Gb
ResNet	3.44Gb	4.82Gb	2.63Gb	1.44Gb
MobileViT	2.78Gb	3.21Gb	2.18Gb	1.23Gb
CoAtPENet	3G	7.87Gb	2.44Gb	1.65Gb

6.6.2 Computational and Communication Costs

In FL, computational and communication efficiency are critical factors that determine the feasibility of deploying models on edge devices with limited resources. This section analyzes the computational complexity, memory requirements, and communication costs of DL models.

Computational Complexity Analysis

Table 6.11 presents the computational complexity in terms of Floating Point Operations (FLOPs) per image for each model. FLOPs provide a hardware-independent measure of the computational work required to process a single image. Among the models evaluated, MobileViT demonstrates the highest computational efficiency with only 1.7G FLOPs, making it particularly suitable for resource-constrained environments. This efficiency comes from its separable convolutions in depth and efficient attention mechanisms specifically designed for mobile applications.

DenseNet follows with 2.9G FLOPs, which benefit from its feature reuse through dense connections that reduce redundant computations. Our proposed CoAtPENet requires 3.4G FLOPs, representing a moderate computational cost with high performance. Although more computationally intensive than MobileViT, CoAtPENet delivers significantly better accuracy (92% vs. 90% on CovidX), justifying the additional computational overhead for applications where accuracy is essential. ResNet50 has

the highest computational demand at 4.1G FLOPs, despite having fewer parameters than some attention-based models, due to its deep convolutional structure.

Memory Requirements

Table 6.12 details the memory consumption during both the training and inference phases. During training, CoAtPENet consumes the highest GPU memory (7.87GB) among the compared models, reflecting the memory-intensive nature of its hybrid architecture, which combines convolutional and attention mechanisms. The increased memory is primarily due to the storage of intermediate activations and gradients during backpropagation. ResNet and DenseNet show moderate GPU memory usage (4.82GB and 6.25GB, respectively), while MobileViT is the most memory-efficient at 3.21GB.

For RAM consumption, ResNet requires the most (3.44GB), followed by CoAtPENet (3GB), MobileViT (2.78GB) and DenseNet (2.58GB). These differences reflect the architecture of each model and the memory needed to store model parameters, optimizer states, and batch data during training.

During inference, memory requirements decrease significantly in all models, as gradient computation and optimization states are no longer needed. CoAtPENet's GPU memory usage drops to 1.65GB, while its RAM consumption decreases to 2.44GB. MobileViT maintains its efficiency advantage with the lowest GPU memory requirement (1.23GB), making it particularly suitable for deployment on devices with limited resources.

All models can function on edge devices such as the NVIDIA Jetson Nano and Raspberry Pi 4 (both with 4GB RAM and 8GB VRAM), as inference requirements are less demanding (maximum 2.7GB RAM and 2GB VRAM). The optimal model choice will depend on specific application needs, including real-time inference requirements or larger batch size handling capabilities.

Communication Costs in Federated Learning

Communication costs are directly proportional to the size of the model, as measured by the number of parameters. From Table 6.11, MobileViT has the smallest size (5.6M parameters), requiring approximately 22.4MB per model transfer (assuming a 32-bit floating-point representation). DenseNet follows with 8M parameters (32MB), while CoAtPENet requires 67.6MB with its 16.9M parameters. ResNet50 has the largest communication overhead at 94MB (23.5M parameters).

For our FL setup with 100 communication rounds and 5 (out of 20) participating

clients per round, the total communication cost would be approximately:

- MobileViT: 22.4 MB $\times$ 5 clients $\times$100 rounds $\approx$ 11.2 GB
- DenseNet: 32 MB $\times$ 5 clients $\times$100 rounds $\approx$ 16 GB
- CoAtPENet: 67.6 MB $\times$ 5 clients $\times$100 rounds $\approx$ 33.8 GB
- ResNet50: 94 MB $\times$ 5 clients $\times$100 rounds $\approx$ 47 GB

These calculations highlight the significant communication overhead in FL, particularly for larger models such as CoAtPENet and ResNet50. However, this analysis also demonstrates the trade-off between model size and performance. Although MobileViT offers the lowest communication cost, CoAtPENet provides better accuracy and robustness in non-IID settings, which may justify the increased communication overhead for applications where performance is critical.

Optimization Strategies

Given the computational and communication challenges identified above, several optimization strategies can be employed to improve efficiency for federated deployment:

- **Model Compression:** Techniques such as pruning (removing redundant parameters) and quantization (reducing parameter precision from 32-bit to 8-bit or lower) can significantly reduce model size. For CoAtPENet, quantization could potentially reduce communication costs by 75% (from 67.6MB to approximately 16.9MB per transfer).
- **Knowledge Distillation:** Training smaller "student" models to mimic the behavior of larger "teacher" models could maintain performance while reducing computational and communication costs.
- **Efficient Communication Protocols:** Implementing techniques like gradient compression, sparsification, or selective parameter updates can substantially reduce communication overhead without significantly affecting model performance.
- **Adaptive Computation:** Dynamically adjusting the computational complexity based on device capabilities and available resources could optimize performance across heterogeneous client devices.

- **Federated Dropout:** Randomly dropping model components during client training can reduce computation while maintaining model diversity and performance.

While CoAtPENet demonstrates better performance in our experiments, particularly in challenging non-IID scenarios, its resource requirements require careful consideration of these optimization strategies for practical deployment. The optimal approach will depend on specific deployment constraints, including device capabilities, network conditions, and accuracy requirements. Future work will explore the implementation of these optimizations to make CoAtPENet more efficient while preserving its performance advantages.

7. Limitations and Future Directions

While our proposed methods demonstrate promising results, a rigorous examination reveals several significant limitations that must be acknowledged and addressed in future work.

7.1 Limitations

FedPANs exhibits concerning sensitivity to hyperparameter selection, particularly period parameters (T) and amplitude (B). As demonstrated in our ablation studies (Figures 5.6, 5.7, and 5.8), minor variations in these parameters can lead to significant performance fluctuations. This sensitivity introduces a substantial challenge for practical deployment, as optimal hyperparameter configurations may vary dramatically across different datasets, model architectures, and client distributions, necessitating extensive and computationally expensive tuning processes for each new application scenario.

While CoAtPENet's hybrid architecture, which integrates convolutional layers, attention mechanisms, and position-aware neurons, introduces a degree of architectural complexity, its parameter count is comparable to, or even lower than, other CNN- or Transformer-based models that achieve similar accuracy. Therefore, the size of the model may not necessarily present a significant limitation. However, the feasibility of deploying CoAtPENet on edge devices with limited computational resources warrants further investigation.

Our implementation of CoAtPENet is highly concentrated on chest X-ray analysis. This specialization raises questions about the generalizability of the model to other medical imaging modalities such as magnetic resonance imaging, CT scans, or ultrasound. The lack of extensive validation in diverse medical imaging domains represents a significant limitation, as it leaves uncertainty as to whether the observed performance advantages would translate to other clinical contexts.

Hybrid models like CoAtPENet can exhibit vulnerability to adversarial attacks, a critical concern in FL environments where malicious clients could potentially compromise the global model. Our work does not adequately address this security dimension, leaving open questions about CoAtPENet's robustness against adversarial examples, model poisoning attacks, or privacy-compromising inference attacks. This security gap represents a significant limitation for deployment in sensitive healthcare settings.

Our custom splitting method to create imbalanced data distributions relies on artificial partitioning of the data sets according to predefined constraints. Although this approach facilitates controlled experimentation, it might not capture the complex organic heterogeneity of the distributions of real-world medical data between institutions. Real clinical data exhibit multidimensional heterogeneity influenced by factors such as patient demographics, institutional specialization, equipment variations, and clinical protocols, nuances that our simplified partitioning approach does not adequately model.

7.2 Future Directions

The hyperparameter sensitivity of FedPANs could be addressed through the development of automated hyperparameter optimization frameworks specifically designed for federated settings. These frameworks could leverage techniques such as Bayesian optimization or evolutionary algorithms to efficiently navigate the hyperparameter space while minimizing communication overhead. Furthermore, meta-learning approaches could be explored to transfer optimal hyperparameter configurations across related tasks, reducing the need for exhaustive tuning in each new application scenario.

Future work should explore systematic architecture simplification strategies for CoAtPENet, potentially leveraging neural architecture search (NAS) techniques to identify more efficient architectures that maintain performance while reducing parameter counts. Knowledge distillation approaches could also be investigated to transfer knowledge from complex CoAtPENet models to simpler and more deployable architectures. These efforts should aim to develop a family of CoAtPENet variants with different complexity-performance trade-offs suitable for diverse deployment scenarios.

To address domain-specific optimization concerns, research should focus on developing cross-modal transfer learning techniques that enable CoAtPENet to generalize across different medical imaging modalities. This could involve pre-training on diverse medical imaging datasets, developing modality-agnostic feature extractors,

or exploring domain adaptation techniques that explicitly account for the unique characteristics of different imaging modalities. Extensive validation across diverse clinical contexts would be essential to establish CoAtPENet's broad applicability.

Inference latency issues on edge devices could be addressed using targeted model compression and acceleration techniques. Approaches such as pruning, quantization, and knowledge distillation could reduce the model size and computational requirements without compromising the performance. Hardware-aware optimization techniques could further improve the efficiency of inference on specific deployment platforms. Furthermore, exploring asynchronous inference pipelines could help mitigate the impact of latency in time-sensitive applications.

To address security concerns, future work should integrate CoAtPENet with robust FL frameworks designed to withstand adversarial attacks. This could involve incorporating techniques such as secure aggregation, differential privacy, Byzantine-robust aggregation, and anomaly detection to identify malicious clients. Formal security analysis and empirical evaluations against state-of-the-art attacks would be essential to establish CoAtPENet's suitability for deployment in security-critical healthcare environments.

Future work should develop more sophisticated data partitioning strategies that better reflect the heterogeneity of medical data in the real world. This could involve analyzing actual multi-institutional medical datasets to characterize authentic patterns of data variation and developing generative models that can simulate these patterns. Furthermore, incorporating domain knowledge about factors that influence data distributions in healthcare settings would enhance the realism of experimental evaluations.

7.3 Potential Applications and Commercialization Opportunities

Despite the limitations discussed above, our proposed methods offer significant potential for practical applications and commercialization in several domains.

CoAtPENet could serve as the basis for privacy-preserving clinical decision support systems that help radiologists diagnose respiratory diseases. By enabling collaborative learning across healthcare institutions, such systems could continuously improve their diagnostic accuracy while complying with regulatory requirements. Commercial products could be developed as software-as-a-service offerings that integrate with existing PACS (Picture Archiving and Communication Systems) and provide real-

time analysis of medical images with appropriate uncertainty quantification.

The proposed FL framework could facilitate the creation of distributed research networks that enable collaborative medical research without centralizing sensitive patient data. Commercial platforms could be developed to manage these networks, providing infrastructure for the training, validation and deployment of secure models across participating institutions. Such platforms could incorporate incentive mechanisms to encourage participation and contribution, potentially through tokenization or other value-sharing approaches.

With increasing regulatory scrutiny of AI in healthcare, our privacy-preserving approach offers a pathway to develop regulatory-compliant AI systems. Commercial services could be established to help AI developers in healthcare use FL for model development and validation while ensuring compliance with regulations such as HIPAA, GDPR, and emerging AI-specific regulations. These services could include compliance certification, audit trails, and documentation generation to streamline the regulatory approval process.

The optimization of CoAtPENet for edge deployment could enable AI-powered diagnostic capabilities in remote or resource-constrained healthcare settings. Commercial products could be developed as specialized hardware devices with embedded AI capabilities that can operate independently while participating periodically in FL to improve their models. Such devices could be particularly valuable in underserved regions with limited access to radiological expertise.

7.4 Summary

While our proposed method demonstrates promising results for FL in medical image analysis, it exhibits significant limitations that must be addressed before widespread practical deployment. Computational overhead and hyperparameter sensitivity limit FedPANs' immediate applicability. The architectural complexity, domain specificity, and potential security vulnerabilities of CoAtPENet present additional challenges.

Addressing these limitations requires concerted research efforts focused on efficiency optimization, theoretical foundation development, cross-modal generalization, security enhancement, and more realistic evaluation frameworks. Despite these challenges, potential applications in clinical decision support, distributed research, regulatory-compliant AI development, and edge healthcare solutions offer compelling commercialization opportunities.

Future work should prioritize not only technical improvements but also interdisciplinary collaboration with healthcare providers, regulatory experts, and ethicists

to ensure that FL solutions for medical imaging are not only technically sound, but also clinically valuable, ethically responsible and practically deployable. By acknowledging and systematically addressing current limitations, we can advance toward FL systems that truly fulfill their promise of enabling collaborative privacy-preserving AI development for improved healthcare outcomes.

8. Conclusion

In this thesis, we have introduced a unified FL framework specifically tailored for medical image classification. Our investigation encompassed a range of DL models, including established architectures such as DenseNet121 and ResNet50, as well as attention-based MobileViT. A key aspect of our research was an in-depth examination of the hybrid CoAtNet model and its suitability for FL scenarios. We thoroughly evaluated these approaches on three benchmark datasets such as CovidX, CheXpert, and MIMIC-CXR, carefully considering multi-label tasks by assigning clients a predetermined number of images per round. Building upon this foundation, we proposed CoAtPENet, a novel integration of CoAtNet with Position-Aware Neurons (PANs) to enhance model performance within the FL setting. By testing CoAtPENet with prominent FL algorithms such as FedAvg and FedProx, we gained valuable insights into its adaptability and robustness under various practical conditions. These included changes in the number of clients, participation ratios, data distributions (IID vs. non-IID, balanced vs. imbalanced) and the activation or deactivation of PANs. Our comprehensive experiments demonstrated that FL algorithms can achieve performance levels comparable to centralized training, underscoring the potential of FL to meet privacy requirements in healthcare settings. By reducing reliance on central data storage, our approach contributes to the ongoing advancement of FL methods for medical applications. Looking ahead, we plan to further refine DL models and FL algorithms that integrate PANs, thereby enhancing the resilience and efficiency of our approach. A promising direction is the development of FL frameworks that can handle multiple vision tasks, such as classification, object detection, and semantic segmentation, within a single system, effectively sharing knowledge between tasks to improve performance and scalability.

9. Appendix

Bibliography

[1] Asmaa Abbas, Mohammed M Abdelsamea and Mohamed Medhat Gaber. '4S-DT: Self-Supervised Super Sample Decomposition for Transfer Learning With Application to COVID-19 Detection'. In: *IEEE Transactions on Neural Networks and Learning Systems* (2021).

[2] Asmaa Abbas, Mohammed M Abdelsamea and Mohamed Medhat Gaber. 'Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network'. In: *Applied Intelligence* 51 (2021), pp. 854–864.

[3] Mohammed Adnan et al. 'Federated learning and differential privacy for medical image analysis'. In: *Scientific reports* 12.1 (2022), p. 1953.

[4] Sakshi Ahuja et al. 'Deep transfer learning-based automated detection of COVID-19 from lung CT scan slices'. In: *Applied Intelligence* 51.1 (2021), pp. 571–585.

[5] Jason Ansel et al. 'PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation'. In: *29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)*. ACM, Apr. 2024. DOI: 10.1145/3620665.3640366. URL: https://pytorch.org/assets/pytorch2-2.pdf.

[6] Mugahed A Al-antari et al. 'Fast deep learning computer-aided diagnosis of COVID-19 based on digital chest x-ray images'. In: *Applied Intelligence* 51.5 (2021), pp. 2890–2907.

[7] Shekoofeh Azizi et al. 'Big self-supervised models advance medical image classification'. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 3478–3488.

[8] Daniel J Beutel et al. 'Flower: A friendly federated learning research framework'. In: *arXiv preprint arXiv:2007.14390* (2020).

[9] David M Blei, Andrew Y Ng and Michael I Jordan. 'Latent dirichlet allocation'. In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022.

[10] Nicolas Carion et al. 'End-to-end object detection with transformers'. In: *European conference on computer vision*. Springer. 2020, pp. 213–229.

[11] Arunava Chakravarty et al. 'Federated learning for site aware chest radiograph screening'. In: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2021, pp. 1077–1081.

[12] Hong-You Chen and Wei-Lun Chao. 'Fedbe: Making bayesian model ensemble applicable to federated learning'. In: *arXiv preprint arXiv:2009.01974* (2020).

[13] Yiqiang Chen et al. 'Fedhealth: A federated transfer learning framework for wearable healthcare'. In: *IEEE Intelligent Systems* 35.4 (2020), pp. 83–93.

[14] Muhammad E. H. Chowdhury et al. 'Can AI Help in Screening Viral and COVID-19 Pneumonia?' In: *IEEE Access* 8 (2020), pp. 132665–132676. ISSN: 2169-3536. DOI: 10.1109/access.2020.3010287. URL: http://dx.doi.org/10.1109/ACCESS.2020.3010287.

[15] Stéphane d'Ascoli et al. 'Convit: Improving vision transformers with soft convolutional inductive biases'. In: *International conference on machine learning*. PMLR. 2021, pp. 2286–2296.

[16] Zihang Dai et al. 'Coatnet: Marrying convolution and attention for all data sizes'. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 3965–3977.

[17] Thomas G Dietterich and Ghulum Bakiri. 'Solving multiclass learning problems via error-correcting output codes'. In: *Journal of artificial intelligence research* 2 (1994), pp. 263–286.

[18] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021. arXiv: 2010.11929 [cs.CV].

[19] Cynthia Dwork, Aaron Roth et al. 'The algorithmic foundations of differential privacy.' In: *Found. Trends Theor. Comput. Sci.* 9.3-4 (2014), pp. 211–407.

[20] Chris Soo-Hyun Eom et al. 'Effective privacy preserving data publishing by vectorization'. In: *Information Sciences* 527 (2020), pp. 311–328.

[21] Alireza Fallah, Aryan Mokhtari and Asuman Ozdaglar. 'Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach'. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 3557–3568.

[22] Ines Feki et al. 'Federated learning for COVID-19 screening from Chest X-ray images'. In: *Applied Soft Computing* 106 (2021), p. 107330.

[23] Chelsea Finn, Pieter Abbeel and Sergey Levine. 'Model-agnostic meta-learning for fast adaptation of deep networks'. In: *International conference on machine learning*. PMLR. 2017, pp. 1126–1135.

[24] 'Forum of International Respiratory Societies. The global impact of respiratory disease'. In: vol. Third. European Respiratory Society. 2021. URL: `firsnet.org/images/publications/FIRS_Master_09202021.pdf`.

[25] Daryl LX Fung et al. 'Self-supervised deep learning model for COVID-19 lung CT image segmentation highlighting putative causal relationship among age, underlying disease and COVID-19'. In: *Journal of Translational Medicine* 19.1 (2021), pp. 1–18.

[26] Johannes Fürnkranz. 'Round robin classification'. In: *Journal of Machine Learning Research* 2.Mar (2002), pp. 721–747.

[27] Navid Ghassemi et al. 'Automatic Diagnosis of COVID-19 from CT Images using CycleGAN and Transfer Learning'. In: *arXiv preprint arXiv:2104.11949* (2021).

[28] Ross Girshick. *Fast R-CNN*. 2015. arXiv: `1504.08083 [cs.CV]`.

[29] Ross Girshick et al. 'Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation'. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2014.

[30] Tripti Goel et al. 'Automatic screening of covid-19 using an optimized generative adversarial network'. In: *Cognitive computation* (2021), pp. 1–16.

[31] Xuan Gong et al. 'Federated learning with privacy-preserving ensemble attention distillation'. In: *IEEE Transactions on Medical Imaging* (2022).

[32] Ian J. Goodfellow et al. *Generative Adversarial Networks*. 2014. arXiv: `1406.2661 [stat.ML]`.

[33] Alex Graves, Abdel-rahman Mohamed and Geoffrey Hinton. 'Speech recognition with deep recurrent neural networks'. In: *2013 IEEE international conference on acoustics, speech and signal processing*. Ieee. 2013, pp. 6645–6649.

[34] Md Jahid Hasan, Md Shahin Alom and Md Shikhar Ali. 'Deep learning based detection and segmentation of COVID-19 & pneumonia on chest X-ray image'. In: *2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*. IEEE. 2021, pp. 210–214.

[35] Kaiming He et al. 'Deep residual learning for image recognition'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[36] Kaiming He et al. *Mask R-CNN*. 2018. arXiv: `1703.06870 [cs.CV]`.

[37] Kaiming He et al. 'Mask r-cnn'. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2961–2969.

[38] Tzu-Ming Harry Hsu, Hang Qi and Matthew Brown. 'Measuring the effects of non-identical data distribution for federated visual classification'. In: *arXiv preprint arXiv:1909.06335* (2019).

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

[40] Shih-Cheng Huang et al. 'Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition'. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 3942–3951.

[41] Jeremy Irvin et al. 'Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison'. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 01. 2019, pp. 590–597.

[42] Md Zabirul Islam, Md Milon Islam and Amanullah Asraf. 'A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images'. In: *Informatics in medicine unlocked* 20 (2020), p. 100412.

[43] Rohan Iyer et al. 'Spatial K-anonymity: A Privacy-preserving Method for COVID-19 Related Geospatial Technologies'. In: *arXiv preprint arXiv:2101.02556* (2021).

[44] Shruti Jadon. 'COVID-19 detection from scarce chest x-ray image data using few-shot deep learning approach'. In: *Medical Imaging 2021: Imaging Informatics for Healthcare, Research, and Applications*. Vol. 11601. International Society for Optics and Photonics. 2021, p. 116010X.

[45] Alistair EW Johnson et al. 'MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports'. In: *Scientific data* 6.1 (2019), p. 317.

[46] Rahma Kadri et al. 'Efficient multimodel method based on transformers and CoAtNet for Alzheimer's diagnosis'. In: *Digital Signal Processing* 143 (2023), p. 104229.

[47] Andrej Karpathy and Li Fei-Fei. *Deep Visual-Semantic Alignments for Generating Image Descriptions*. 2015. arXiv: `1412.2306 [cs.CV]`.

[48] Diederik P Kingma and Jimmy Ba. 'Adam: A method for stochastic optimization. iclr. 2015'. In: 9 (2015).

[49] Alexander Kirillov et al. 'Panoptic segmentation'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 9404–9413.

[50] Jakub Konečný et al. 'Federated Learning: Strategies for Improving Communication Efficiency'. In: *NIPS Workshop on Private Multi-Party Machine Learning*. 2016. URL: `https://arxiv.org/abs/1610.05492`.

[51] Pranav Kulkarni et al. 'Optimizing Federated Learning for Medical Image Classification on Distributed Non-iid Datasets with Partial Labels'. In: *arXiv preprint arXiv:2303.06180* (2023).

[52] Rajesh Kumar et al. 'Blockchain-federated-learning and deep learning models for covid-19 detection using ct imaging'. In: *IEEE Sensors Journal* 21.14 (2021), pp. 16301–16314.

[53] Daliang Li and Junpu Wang. 'Fedmd: Heterogenous federated learning via model distillation'. In: *arXiv preprint arXiv:1910.03581* (2019).

[54] Kunchang Li et al. 'Uniformer: Unifying convolution and self-attention for visual recognition'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.10 (2023), pp. 12581–12600.

[55] Tian Li et al. 'Federated optimization in heterogeneous networks'. In: *Proceedings of Machine Learning and Systems* 2 (2020), pp. 429–450.

[56] Wei Li et al. 'NIA-Network: Towards improving lung CT infection detection for COVID-19 diagnosis'. In: *Artificial Intelligence in Medicine* 117 (2021), p. 102082.

[57] Weisheng Li et al. 'DFENet: A dual-branch feature enhanced network integrating transformers and convolutional feature learning for multimodal medical image fusion'. In: *Biomedical Signal Processing and Control* 80 (2023), p. 104402.

[58] Xiaoxiao Li et al. 'Fedbn: Federated learning on non-iid features via local batch normalization'. In: *arXiv preprint arXiv:2102.07623* (2021).

[59] Xin-Chun Li et al. 'Federated learning with position-aware neurons'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 10082–10091.

[60] Zekun Li et al. 'A novel multiple instance learning framework for COVID-19 severity assessment via data augmentation and self-supervised learning'. In: *Medical Image Analysis* 69 (2021), p. 101978.

[61] Zonggui Li et al. 'COVID-19 Diagnosis on CT Scan Images Using a Generative Adversarial Network and Concatenated Feature Pyramid Network with an Attention Mechanism'. In: *Medical Physics* (2021).

[62] Tao Lin et al. 'Ensemble distillation for robust model fusion in federated learning'. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 2351–2363.

[63] Boyi Liu et al. 'Experiments of federated learning for covid-19 chest x-ray images'. In: *arXiv preprint arXiv:2007.05592* (2020).

[64] Shu Liu et al. 'Path aggregation network for instance segmentation'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 8759–8768.

[65] Jun Ma et al. 'Toward data-efficient learning: A benchmark for COVID-19 CT lung and infection segmentation'. In: *Medical physics* 48.3 (2021), pp. 1197–1210.

[66] Nursultan Makhanov, Nguyen Anh Tu and Kok-Seng Wong. 'A Survey on Deep Learning Advances and Emerging Issues in Pneumonia and COVID19 Prediction'. In: *2022 IEEE International Conference on Big Data and Smart Computing (BigComp)*. IEEE. 2022, pp. 96–103.

[67] Disha Makhija et al. 'Architecture agnostic federated learning for neural networks'. In: *International Conference on Machine Learning*. PMLR. 2022, pp. 14860–14870.

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

[69] Sachin Mehta and Mohammad Rastegari. 'MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer'. In: *International Conference on Learning Representations*. 2022. URL: https://openreview.net/forum?id=vh-0sUt8HlG.

[70] Zümrüt Müftüoğlu, M Ayyüce Kizrak and Tülay Yildlnm. 'Differential Privacy Practice on Diagnosis of COVID-19 Radiology Imaging Using EfficientNet'. In: *2020 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*. IEEE. 2020, pp. 1–6.

[71] Hemalatha Munusamy et al. 'FractalCovNet architecture for COVID-19 Chest X-ray image Classification and CT-scan image Segmentation'. In: *Biocybernetics and Biomedical Engineering* 41.3 (2021), pp. 1025–1038.

[72] A. Nayyar, R. Jain and Y. Upadhyay. 'Object detection based approach for Automatic detection of Pneumonia'. In: *2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*. 2020, pp. 1–6.

[73] Sadaf Naz, Khoa T Phan and Yi-Ping Phoebe Chen. 'A comprehensive review of federated learning for COVID-19 detection'. In: *International Journal of Intelligent Systems* 37.3 (2022), pp. 2371–2392.

[74] Yuto Nishitaki, Tohru Kamiya and Shoji Kido. 'Identification of Nodular Shadows from CT Images Using Improved CoAtNet Incorporated Clinical Recording'. In: *2023 23rd International Conference on Control, Automation and Systems (ICCAS)*. IEEE. 2023, pp. 1727–1732.

[75] Hyeonwoo Noh, Seunghoon Hong and Bohyung Han. 'Learning Deconvolution Network for Semantic Segmentation'. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Dec. 2015.

[76] Jacquelyn K O'herrin, Norman Fost and Kenneth A Kudsk. 'Health Insurance Portability Accountability Act (HIPAA) regulations: effect on medical record research'. In: *Annals of surgery* 239.6 (2004), p. 772.

[77] Junghoon Park, Il-Youp Kwak and Changwon Lim. 'A Deep Learning Model with Self-Supervised Learning and Attention Mechanism for COVID-19 Diagnosis Using Chest X-ray Images'. In: *Electronics* 10.16 (2021), p. 1996.

[78] Sangjoon Park et al. 'Federated Split Task-Agnostic Vision Transformer for COVID-19 CXR Diagnosis'. In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., 2021, pp. 24617–24630. URL: https://proceedings.neurips.cc/paper/2021/file/ceb0595112db2513b9325a85761b7310-Paper.pdf.

[79] Sangjoon Park et al. *Vision Transformer for COVID-19 CXR Diagnosis using Chest X-ray Feature Corpus*. 2021. arXiv: 2103.07055 [eess.IV].

[80] Tuan D Pham. 'A comprehensive study on classification of COVID-19 on computed tomography with pretrained convolutional neural networks'. In: *Scientific reports* 10.1 (2020), pp. 1–8.

[81] Tuan D. Pham. 'A comprehensive study on classification of COVID-19 on computed tomography with pretrained convolutional neural networks'. In: *Scientific Reports* 10.1 (2020), p. 16942. DOI: 10.1038/s41598-020-74164-z. URL: https://doi.org/10.1038/s41598-020-74164-z.

[82] Daniel Povey et al. 'The Kaldi speech recognition toolkit'. In: *IEEE 2011 workshop on automatic speech recognition and understanding*. CONF. IEEE Signal Processing Society. 2011.

[83] KV Priya and J Dinesh Peter. 'A federated approach for detecting the chest diseases using DenseNet for multi-label classification'. In: *Complex & Intelligent Systems* (2021), pp. 1–9.

[84] Denis A Pustokhin et al. 'An effective deep residual network based class attention layer with bidirectional LSTM for diagnosis and classification of COVID-19'. In: *Journal of Applied Statistics* (2020), pp. 1–18.

[85] Tran Minh Quan et al. 'XPGAN: X-Ray Projected Generative Adversarial Network For Improving Covid-19 Image Classification'. In: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2021, pp. 1509–1513.

[86] Rishi Raj et al. 'StrokeViT with AutoML for brain stroke classification'. In: *Engineering Applications of Artificial Intelligence* 119 (2023), p. 105772.

[87] Vignav Ramesh, Blaine Rister and Daniel L Rubin. 'COVID-19 Lung Lesion Segmentation Using a Sparsely Supervised Mask R-CNN on Chest X-rays Automatically Computed from Volumetric CTs'. In: *arXiv preprint arXiv:2105.08147* (2021).

[88] Jesse Read et al. 'Classifier chains for multi-label classification'. In: *Machine learning* 85 (2011), pp. 333–359.

[89] Sashank J Reddi et al. 'Adaptive Federated Optimization'. In: *International Conference on Learning Representations*.

[90] Joseph Redmon et al. *You Only Look Once: Unified, Real-Time Object Detection*. 2016. arXiv: 1506.02640 [cs.CV].

[91] Joseph Redmon et al. 'You only look once: Unified, real-time object detection'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 779–788.

[92] Shaoqing Ren et al. *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks*. 2016. arXiv: 1506.01497 [cs.CV].

[93] Ryan Rifkin and Aldebaro Klautau. 'In defense of one-vs-all classification'. In: *Journal of machine learning research* 5.Jan (2004), pp. 101–141.

[94] Olaf Ronneberger, Philipp Fischer and Thomas Brox. 'U-net: Convolutional networks for biomedical image segmentation'. In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, pp. 234–241.

[95] Mark Sandler et al. 'Mobilenetv2: Inverted residuals and linear bottlenecks'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 4510–4520.

[96] Ahmed Sedik et al. 'Efficient deep learning approach for augmented detection of Coronavirus disease'. In: *Neural Computing and Applications* (2021), pp. 1–18.

[97] Debaditya Shome et al. 'Covid-transformer: Interpretable covid-19 detection using vision transformer for healthcare'. In: *International Journal of Environmental Research and Public Health* 18.21 (2021), p. 11086.

[98] Mohammad Shorfuzzaman and M Shamim Hossain. 'MetaCOVID: A Siamese neural network framework with contrastive loss for n-shot diagnosis of COVID-19 patients'. In: *Pattern recognition* 113 (2021), p. 107700.

[99] Sidak Pal Singh and Martin Jaggi. 'Model fusion via optimal transport'. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 22045–22055.

[100] Anuroop Sriram et al. 'COVID-19 Prognosis via Self-Supervised Representation Learning and Multi-Image Prediction'. In: *arXiv preprint arXiv:2101.04909* (2021).

[101] Canh T Dinh, Nguyen Tran and Josh Nguyen. 'Personalized federated learning with moreau envelopes'. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 21394–21405.

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

[103] Shanjiang Tang et al. 'EDL-COVID: Ensemble Deep Learning for COVID-19 Cases Detection from Chest X-Ray Images'. In: *IEEE Transactions on Industrial Informatics* (2021).

[104] Colin Tankard. 'What the GDPR means for businesses'. In: *Network Security* 2016.6 (2016), pp. 5–8.

[105] Soroosh Tayebi Arasteh et al. 'Enhancing domain generalization in the AI-based analysis of chest radiographs with federated learning'. In: *Scientific Reports* 13.1 (2023), p. 22576.

[106] Pun Liang Thon et al. 'Investigation of ConViT on COVID-19 Lung Image Classification and the Effects of Image Resolution and Number of Attention Heads'. In: *International Journal of Integrated Engineering* 15.3 (2023), pp. 54–63.

[107] Grigorios Tsoumakas and Ioannis Katakis. 'Multi-label classification: An overview'. In: *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications* (2008), pp. 64–74.

[108] Jasper RR Uijlings et al. 'Selective search for object recognition'. In: *International journal of computer vision* 104.2 (2013), pp. 154–171.

[109] Anwaar Ulhaq and Oliver Burmeister. *COVID-19 Imaging Data Privacy by Federated Learning Design: A Theoretical Framework*. 2020. arXiv: `2010.06177 [cs.CR]`.

[110] Ashish Vaswani et al. 'Attention is all you need'. In: *Advances in neural information processing systems* 30 (2017).

[111] Hongyi Wang et al. 'Federated Learning with Matched Averaging'. In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL: `https://openreview.net/forum?id=BkluqlSFDS`.

[112] Hongyi Wang et al. 'Federated Learning with Matched Averaging'. In: *International Conference on Learning Representations*. 2020.

[113] Linda Wang, Zhong Qiu Lin and Alexander Wong. 'Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images'. In: *Scientific Reports* 10.1 (2020), pp. 1–12.

[114] Linda Wang and Alexander Wong. *COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest X-Ray Images*. 2020. arXiv: 2003.09871 [eess.IV].

[115] Haiping Wu et al. 'Cvt: Introducing convolutions to vision transformers'. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 22–31.

[116] Yu-Huan Wu et al. *JCS: An Explainable COVID-19 Diagnosis System by Joint Classification and Segmentation*. 2020. arXiv: 2004.07054 [eess.IV].

[117] Pooja Yadav et al. 'Lung-GANs: Unsupervised Representation Learning for Lung Disease Classification Using Chest CT and X-Ray Images'. In: *IEEE Transactions on Engineering Management* (2021).

[118] Samir S Yadav and Shivajirao M Jadhav. 'Deep convolutional neural network based medical image classification for disease diagnosis'. In: *Journal of Big data* 6.1 (2019), pp. 1–18.

[119] Hongwei Yang et al. 'FedSteg: A federated transfer learning framework for secure image steganalysis'. In: *IEEE Transactions on Network Science and Engineering* 8.2 (2020), pp. 1084–1094.

[120] Qian Yang et al. 'Flop: Federated learning on medical datasets using partial networks'. In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2021, pp. 3845–3853.

[121] Seongjun Yang et al. 'Towards the Practical Utility of Federated Learning in the Medical Domain'. In: *arXiv preprint arXiv:2207.03075* (2022).

[122] Zong-Ye Yang and Qiangfu Zhao. 'A Multiple Deep Learner Approach for X-Ray Image-Based Pneumonia Detection'. In: *2020 International Conference on Machine Learning and Cybernetics (ICMLC)*. IEEE. 2020, pp. 70–75.

[123] Shangjie Yao et al. 'Pneumonia Detection Using an Improved Algorithm Based on Faster R-CNN'. In: *Computational and Mathematical Methods in Medicine* 2021 (2021).

[124] Fuxun Yu et al. 'Fed2: Feature-aligned federated learning'. In: *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*. 2021, pp. 2066–2074.

[125] Danni Yuan et al. 'Collaborative deep learning for medical image analysis with differential privacy'. In: *2019 IEEE Global Communications Conference (GLOBECOM)*. IEEE. 2019, pp. 1–6.

[126] Mikhail Yurochkin et al. 'Bayesian nonparametric federated learning of neural networks'. In: *International Conference on Machine Learning*. PMLR. 2019, pp. 7252–7261.

[127] Mikhail Yurochkin et al. 'Statistical model aggregation via parameter matching'. In: *Advances in neural information processing systems* 32 (2019).

[128] Manzil Zaheer et al. 'Adaptive methods for nonconvex optimization'. In: *Advances in neural information processing systems* 31 (2018).

[129] Ju Zhang et al. 'Dense GAN and Multi-layer Attention based Lesion Segmentation Method for COVID-19 CT Images'. In: *Biomedical Signal Processing and Control* (2021), p. 102901.

[130] Min-Ling Zhang and Zhi-Hua Zhou. 'A review on multi-label learning algorithms'. In: *IEEE transactions on knowledge and data engineering* 26.8 (2013), pp. 1819–1837.

[131] Weishan Zhang et al. 'Dynamic fusion-based federated learning for COVID-19 detection'. In: *IEEE Internet of Things Journal* (2021).

[132] Tongxue Zhou, Stéphane Canu and Su Ruan. 'Automatic COVID-19 CT segmentation using U-Net integrated spatial and channel attention mechanism'. In: *International Journal of Imaging Systems and Technology* 31.1 (2021), pp. 16–27.