

Towards Multi-Task Medical Imaging Models: Exploring Federated Learning with Vision Transformers

MS (Research) Thesis

By
Anirban Nath



**DISCIPLINE OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY INDORE**

April 2025

Towards Multi-Task Medical Imaging Models: Exploring Federated Learning with Vision Transformers

A THESIS

*Submitted in fulfillment of the
requirements for the award of the degree
of*

Master of Science (Research)

by

Anirban Nath



**DISCIPLINE OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY INDORE**

April 2025



INDIAN INSTITUTE OF TECHNOLOGY INDORE

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the thesis entitled **Towards Multi-Task Medical Imaging Models: Exploring Federated Learning with Vision Transformers** in the fulfillment of the requirements for the award of the degree of **MASTER OF SCIENCE (RESEARCH)** and submitted in the **DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, Indian Institute of Technology Indore**, is an authentic record of my own work carried out during the time period from July 2021 to April 2025 under the supervision of Dr. Puneet Gupta, Associate Professor, Department of Computer Science and Engineering.

The matter presented in this thesis has not been submitted by me for the award of any other degree of this or any other institute.

Anirban Nath
23/04/25

Signature of the student with date
(Anirban Nath)

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

(24/04/2025)

Signature of the Supervisor of
MS (Research) thesis (with date)
(Dr. Puneet Gupta)

Anirban Nath has successfully given his/her MS (Research) Oral Examination held on

07 July 2025

07/07/2025
Signature of Chairperson (OEB) with date

07/07/2025
Signature(s) of Thesis Supervisor(s) with date

07/07/25
Signature of Convener, DPGC with date

07/07/25
Signature of Head of Discipline with date

ACKNOWLEDGEMENTS

I would like to take this opportunity to thank everyone whose efforts have made this project successful. First and foremost, I would like to express my deepest regard for my thesis supervisor, **Dr. Puneet Gupta**, who has guided me through every step of my thesis journey. It is only because of his guidance and vision that I was able to work on such an interesting topic and face all challenges that came in the due course of my degree. His constant support and encouragement kept me motivated throughout the course of my work.

I would also like to express my gratitude to **Dr. Ranveer Singh**, Head of Department, and **Prof. Somnath Dey**, former Head of Department, Computer Science and Engineering, for all their help and support. I am also extremely thankful to **Dr. Surya Prakash** and **Dr. Girish Verma**, my research progress committee members, for taking the time to periodically evaluate my thesis progress and providing valuable feedback throughout.

My respect to **Prof. Suhas Joshi**, Director, Indian Institute of Technology Indore, for providing me with the opportunity to explore my research capabilities at the Indian Institute of Technology Indore.

A sincere token of thanks to my seniors **Mr. Anup Kumar Gupta** and **Ms. Sneha Shukla**, whose guidance and constant support made this journey productive and enjoyable. I would also like to thank my friends **Mr. Aditya Dixit**, and my juniors **Mr. Ashutosh Dhamaniya** and **Mr. Nischit Hosamani** for this beautiful experience.

I thank **Ms. Geena Anjelus Samar** for her constant support throughout this journey. I thank my parents for believing in me and supporting me at all times. They provided unwavering support at each and every step of my life. I thank my grandparents, uncle, aunt and my brothers and sisters for their love. Finally, I would like to express my thanks to God almighty for providing me with great and supportive people throughout.

Anirban Nath

Dedicated to my parents

मेरे माता-पिता को समर्पित

Abstract

Medical Imaging models have become commonplace for critical diagnostic tasks such as image segmentation, detection, and classification. They have been proven to perform better than humans and have made diagnostic procedures largely hassle-free with minimum human intervention. However, the training of robust diagnostic models is hindered by two major roadblocks. Firstly, training specialized models for each task requires large amounts of data. Secondly, several privacy laws restrict the sharing of medical data, limiting opportunities for collaborative training. To overcome the first challenge, Multi-Task Learning (MTL) is utilized to perform multiple tasks using a single model. However, while traditional Convolutional Neural Network-based MTL models excel at identifying local features, they struggle to contextualize global features. To address the second challenge, Federated Learning (FL) is used to collaboratively train models by periodically sharing model weights with an aggregation server, avoiding direct data communication. However, neural networks are permutation invariant, which means that permuting the nodes in any layer of the network does not affect its prediction outcome. This is a problem for traditional FL methods, as averaging the weight tensors of corresponding layers of multiple separately trained models can result in distorted feature maps. To address these concerns, we propose DiagnoFormer, a transformer-based multi-task model that exploits the task-agnostic feature learning capability of transformers to learn robust features for the aforementioned imaging tasks using a common encoder and task-specific decoders. By combining MTL with a hybrid loss function, DiagnoFormer learns distinct diagnostic tasks in a synergistic manner. We also introduce a novel Bayesian Federation algorithm for multitask imaging models, which circumvents the need for direct parameter averaging while also enabling simultaneous segmentation and classification. Experiments on publicly available mammogram and pneumonia datasets demonstrate that DiagnoFormer outperforms popular single-task and MTL models. Furthermore, our Bayesian Federation method surpasses traditional FL methods for image segmentation.

Publications

1. **Nath A**, Shukla S, Gupta P (**Accepted**) MTMedFormer: Multi-Task Vision Transformer for Medical Imaging with Federated Learning. Medical & Biological Engineering & Computing. (2025) Impact Factor- 2.6. DOI: 10.1007/s11517-025-03404-z

Table of Contents

Abstract	i
List of Figures	v
List of Tables	vii
Nomenclature	viii
Acronyms	x
1 Introduction	1
1.1 Thesis Contribution	3
1.2 Thesis Organization	4
2 Literature Survey	5
2.1 Multi-Task Learning in Medical Imaging	5
2.2 Federated Learning in Medicine	6
3 Methodology	8
3.1 Proposed Model: DiagnoFormer	9
3.1.1 Task-Agnostic Encoder	10
3.1.2 Task-Specific Decoders	11
3.2 Loss Functions	12
3.2.1 Task Specific Losses	12
3.2.2 Hybrid Loss	12

3.3	Federation Schemes	12
3.3.1	Federated Averaging (FedAvg)	13
3.3.2	Partial Participation Federation (PPF)	13
3.3.3	Post-federation Finetuning	15
3.3.4	Federated Bayesian Ensemble (FBE)	15
4	Experimental Evaluation	18
4.1	Dataset and Metrics	18
4.2	Organization of Federation	20
4.3	Dataset Augmentation	21
4.4	Choice of Training Parameters	22
4.5	Comparative Evaluation	22
4.5.1	Performance Analysis of DiagnoFormer	22
4.5.2	Analysis of Federation Methods	24
4.5.3	Model Sizes and Communication Benefits	26
4.6	Ablation Studies	27
4.6.1	Replacing Transformer Encoder with CNN UNet	27
4.6.2	Analysis of Differential Privacy	29
5	Conclusion and Future Scope	32

List of Figures

3.1	Architecture of DiagnoFormer. The encoder comprises multi-headed self-attention layers followed by a depthwise convolution layer (DC). The Encoder Stage 4 output is sent to all the task-specific decoders for generating respective outputs. The hybrid loss is calculated by taking a weighted sum of the task-specific losses from the decoders. (H, W, C) refer to height, width, and number of channels respectively.	9
3.2	Overview of FedAvg. Given N clients, the server performs federated averaging and sends back w_{avg} to clients. Here M_i and w_i refer to client models and their respective locally trained weights. $ D_i $ denotes the dataset size of client i and $ D = \Sigma D_i $	13
3.3	Overview of Partial Participation Federation. Encoder weights w_{avg}^{enc} and decoder weights w_{avg}^{task} are computed separately. M_i and w_i refer to client models and their respective locally trained weights, whereas D refers to the client dataset.	14
3.4	Overview of Federated Bayesian Ensembling. The client models upload their weights to the server where a Gaussian distribution is fitted to them. Teacher models are sampled from this distribution and used to train the server model through knowledge distillation. M_i are the client models, U refer to unlabeled data samples, and $\hat{y}_{teacher}$ refer to the pseudo-labels output by the sampled teacher models.	16
4.1	Qualitative results of DiagnoFormer on INBreast dataset	24
4.2	Qualitative results of DiagnoFormer on v7-Labs Darwin dataset	25

4.3	Overview of DP-SGD with $\varepsilon = 20$ and $\varepsilon = 47$. X-Axis: Federation Round, Y-Axis: Metric (i) Segmentation (Dice Score) (ii) Detection (mAP) (iii) Classification (AUC).	30
-----	---	----

List of Tables

4.1	Dataset division of v7-Labs dataset for all federated training and evaluation. PNA stands for Pneumonia. Kindly note that only training images are augmented.	20
4.2	Augmentations applied to training dataset. H. Flip is Horizontal Flip. ✓/✗represents instances where the augmentation has been randomly applied half the time.	21
4.3	Performance of DiagnoFormer against existing models.	23
4.4	Performance of DiagnoFormer under data centralized, and various federated and finetuned settings.	26
4.5	Model Sizes and Parameter Count Comparison.	27
4.6	Performance variation between clients using the CNN-based model vs DiagnoFormer.	28

Nomenclature

H	Height of image
W	Width of image
C	Number of feature channels
E	Encoder
B	Decoder
D	Dataset
p	X-coordinate of a given pixel
q	Y-coordinate of a given pixel
α	Receptive field width in convolution operation
β	Receptive field height in convolution operation
K	Convolution Kernel
l	Loss function
σ	Weight hyperparameter for task-specific losses
$mask$	Segmentation Mask
$label$	Classification Label
I	Object Query

th	Confidence Threshold
U	Unlabeled data
Q	Matrix containing the Queries
K	Matrix containing the Keys
V	Matrix containing the Values
M	Client Model
T	Teacher Model
w	Weight matrix
SA	Self Attention
\hat{y}	Prediction
y	Ground Truth

Acronyms

AUC	Area Under Curve
CE	Cross Entropy
CNN	Convolutional Neural Network
CV	Computer Vision
DC	Depth Wise Convolution Layer
DiagnoFormer	Diagnostic Multi-Task Medical Vision TransFormer
DETR	Detection Transformer
DL	Deep Learning
DNN	Deep Neural Network
DP	Differential Privacy
ECG	Electrocardiogram
FBE	Federated Bayesian Ensemble
FedAvg	Federated Averaging
FL	Federated Learning
GELU	Gaussian Error Linear Unit,
gIoU	Generalized Intersection over Union

IoT	Internet of Things
MLP	Multilayer Perceptron
MHA	Multi-headed Self Attention
MTL	Multi-Task Learning
NLP	Natural Language Processing
PLD	Progressive Locality Decoder
PNA	Pneumonia
PPF	Partial Participation Federation
PVT	Pyramidal Vision Transformer
ResNet	Residual Neural Network
SGD	Stochastic Gradient Descent
SWA	Stochastic Weighted Averaging
TPR	True Positive Rate
TPI	False Positive Rate per Image
ViT	Vision Transformer

Chapter 1

Introduction

Deep Learning (DL) in medical imaging is a constantly evolving research discipline at the intersection of medicine and artificial intelligence. Particularly, its applications in precision medicine has emerged as a significant area of research (Santos et al., 2019). Precision medicine is the practice of tailoring treatments based on individual symptom variations. This is crucial for facilitating targeted interventions. In that regard, medical imaging techniques have significantly enhanced medical practitioners' ability to swiftly and accurately pinpoint problematic areas within scan images.

A key requirement for precision medicine is precision diagnostics (Raparthi et al., 2021). In medical imaging, precision diagnostics comprises three fundamental tasks: image segmentation, object detection, and image classification. Segmentation involves generating pixel-level binary masks to precisely delineate diseased regions. Object detection entails the identification of regions of interest by enclosing them within bounding boxes and providing precise details about their dimensions and centre points. Image classification is the process of identifying the type or nature of anomaly present within a given image. Although several existing models have been developed for each of the aforementioned imaging tasks (Carion et al., 2020, Dosovitskiy et al., 2020, Zhang et al., 2021), they offer only partial information about the detected abnormalities. For example, segmentation only highlights the pixels belonging to the affected region, while classification predicts just the type of disease. For complex diagnostic tasks, such as cancer detection, details like the size, centre, shape, and class of detected

tumours are all crucial for a precise diagnosis (Gurcan et al., 2009).

To enable models to perform more than one task through a singular architecture, Multi-Task Learning (MTL) is often employed. In MTL, dedicated sub-models are designed to handle specific tasks within a unified model framework. These sub-models typically share a common loss function, which is used to optimize the overall model. However, most MTL models perform best when the tasks being performed require very similar features. When distinct features need to be extracted per task, traditional MTL models become infeasible (Graham et al., 2023). To address this, a few MTL models have been recently proposed that share a portion of their architecture (Gao et al., 2020, Park et al., 2021). Such models are called pipelined MTL models. A pipelined approach to MTL models allows for the learning of task-agnostic feature maps. This mitigates the need for separate retraining or fine-tuning of models for individual tasks, particularly when data is limited. Pipelined MTL models also typically have fewer trainable parameters (Lu et al., 2020), which makes them easier to train and deploy in resource-constrained environments.

However, most pipelined MTL imaging models proposed thus far rely on a CNN backbone architecture (Gao et al., 2020, Graham et al., 2023, Zhou et al., 2021), such as ResNet or UNet. In contrast, Vision Transformers (Dosovitskiy et al., 2020) exhibit greater resilience to visual disturbances (Filipiuk and Singh, 2022). This is particularly crucial for medical imaging where certain complex biological structures within scan images may not be relevant to the identification task. Such perturbations can lead to issues like false positives in detection or imprecise segmentation.

A reduction in computational cost also holds significant potential for distributed model training. In this context, Federated Learning (McMahan et al., 2017) (FL) has emerged as a promising research direction, with diverse applications in distributed healthcare. FL has gained particular relevance with the rise of Internet of Things (IoT)-enabled healthcare devices (Chakraborty et al., 2023). Coupled with rapid advancements in edge computing in recent years (Ghosh and Ghosh, 2023), IoT devices have become powerful enough for brief local training of DL models. However, only a limited number of federated MTL medical imaging models have been proposed (Park

et al., 2021), while most existing approaches remain either single-task or CNN-based (Dayan et al., 2021, Sheller et al., 2020, Zhang et al., 2020). As a result, the performance and structure of these models leave room for substantial improvement. Furthermore, traditional FL methods face training-time challenges, such as permutation invariance and model divergence (Karimireddy et al., 2020), which may hinder their effectiveness for highly parameterized models.

Motivated by the aforementioned research gaps, we propose a novel **Diagnostic Multi-Task Medical Vision TransFormer** model, DiagnoFormer, designed for simultaneous image segmentation, detection, and classification. Our model features a task-agnostic pyramidal vision transformer encoder paired with task-specific decoders, enabling concurrent execution of three imaging tasks on a single input image. Additionally, we introduce a Bayesian learning-based FL technique called Federated Bayesian Ensemble (FBE) for multi-task imaging models that mitigates the pitfalls of traditional FL methods. To validate our approach, we conduct extensive experiments to evaluate our model’s performance against popular single-task and multi-task models. We also assess DiagnoFormer’s efficacy under various federation methods, including FBE.

1.1 Thesis Contribution

The main contributions of the thesis are as follows:-

1. We introduce DiagnoFormer, a novel Transformer-based MTL imaging network that unifies multiple imaging tasks through a shared encoder pipeline fitted with task-specific decoders. Our model effectively learns robust task-agnostic features to perform three crucial diagnostic imaging tasks: image segmentation, object detection, and image classification on a single input image.
2. We propose Federated Bayesian Ensemble (FBE), a novel Bayesian federated learning technique for multi-task image segmentation and classification. To the best of our knowledge, we are the first to demonstrate the application of Bayesian

Federation for multi-task Transformer-based models for complex imaging tasks. Furthermore, we conduct a comprehensive analysis of DiagnoFormer under various other federation techniques.

3. Experiments using multiple disease identification datasets demonstrate that DiagnoFormer outperforms existing multi-task models and matches or exceeds the performance of popular single-task models. We also show that FBE outperforms other federation methods for segmentation.

1.2 Thesis Organization

The rest of the thesis is organized in the following manner. In Chapter 2, we provide a brief overview of the relevant works on MTL models and the usage of FL in medical imaging. In Chapter 3, we present the proposed DiagnoFormer model and detail the working of its various components. We also discuss the FL techniques evaluated in this paper and introduce our proposed Bayesian Federation method, FBE. In Chapter 4, we outline the datasets used, experimental settings, and discuss our results and findings. In Chapter 5, we conclude by summarising our work and briefly discussing potential future research directions.

Chapter 2

Literature Survey

In this chapter, we delineate the existing research into multi-task medical imaging models and the usage of federated learning for distributed model training within medicine.

2.1 Multi-Task Learning in Medical Imaging

Multi-Task Learning (MTL) models leverage the interdependence of related tasks to produce multiple outputs using a single model. These models typically consist of shared layers that learn task-agnostic features, which are then passed to task-specific heads for final prediction (Zhou et al., 2021). Some early works include that of Akselrod-Ballin et al. (2016), in which the authors proposed a multi-task ResNet-based model for simultaneous detection and classification of tumours. Their model comprised task-specific heads that utilized feature maps extracted by the ResNet backbone. Similarly, Zhou et al. (2021) implemented a 3D UNet architecture for breast cancer segmentation and classification. Segmentation was performed by the main UNet network, while a classification output was extracted through convolution layers applied to feature maps extracted from the bridge block. Graham et al. (2023) proposed an MTL model with a ResNet backbone and type-specific segmentation heads for multi-label segmentation. Gao et al. (2020) used a ResNet50-based backbone as a shared feature encoder, integrating task-specific heads for segmentation, detection,

and classification.

While the aforementioned models highlight the effectiveness of multi-task architectures, they all rely on CNN backbones. While fast and generally reliable, CNNs have been shown to struggle with capturing long-range dependencies effectively (Li et al., 2022). Vision Transformers (ViTs) (Dosovitskiy et al., 2020) address this limitation by dividing images into non-overlapping patches and finding dependencies across them. Chen et al. (2021) demonstrated that transformer-based feature representations are robust enough to be shared across tasks. Their model utilized a shared encoder-decoder framework with task-specific heads and tails for image denoising and super-resolution. Similarly, Kim et al. (2022) employed a task-agnostic transformer model as a common feature extractor, enabling a plug-and-play approach for various imaging tasks using task-specific decoders. In medical imaging, a similar approach was proposed by Park et al. (2021). The authors introduced FeSTA, a federated split multi-task model that could be used to perform medical image segmentation, detection, or classification, based on client needs. However, these transformer models struggle to generalise effectively across all tasks and demand extensive training. Notably, FeSTA requires thousands of training epochs to achieve competitive performance.

2.2 Federated Learning in Medicine

Federated Learning (McMahan et al., 2017) (FL) is a distributed model training paradigm in which a global model is trained by aggregating model weights from multiple client models. Clients in discrete locations train individual models on their local data and periodically send model weights to a central server for aggregation. This eliminates the need to transmit sensitive data to a centralized location, which makes the training process secure. FL has emerged as a promising research direction for collaborative medical AI, as it enables the development of robust models while preserving patient data privacy.

Early research into FL for medical taskflows includes the work of Jochems et al. (2017). The authors used FL to train a global Bayesian network for predicting the

two-year survival probability of lung cancer patients. Sheller et al. (2020) distributed the BRATS 2017 dataset (Menze et al., 2014) across multiple clients and evaluated FL methods against incremental learning and non-federated data-sharing approaches. Their findings showed that models aggregated with FL were stable and performed on par with alternative learning methods. Another large-scale study by Dayan et al. (2021) conducted during the COVID-19 pandemic demonstrated that global models produced using FL outperformed client models trained on limited datasets.

The application of FL techniques in medicine has gained significance with the widespread adoption of medical IoT devices. In this context, Zhang et al. (2020) proposed an FL-based model for analyzing electrocardiogram (ECG) data to detect arrhythmia using data collected through medical IoT devices. Similarly, Can and Ersoy (2021) explored the usage of FL for ECG data analysis to predict stress levels of test subjects in a privacy-focused manner.

None of these models, however, addresses the pitfalls of traditional FL methods such as model divergence and permutation invariance. Additionally, none of these models are built on the transformer architecture, which limits their performance for complex diagnostic tasks. In fact, FeSTA (Park et al., 2021) is the only notable model proposed for federated multi-task medical imaging that utilizes transformers. FeSTA is designed with server-side (task-agnostic) and client-side (task-specific) components. The task-specific components are periodically federated among participating clients. However, FeSTA’s architecture requires frequent transmission of features between the server and clients for each training batch, leading to high bandwidth consumption.

Chapter 3

Methodology

Precision Diagnostics involves comprehensive localization of anomalies to administer the best course of treatment. Task-specific models provide partial information on the nature of the anomaly, requiring further examination to fully understand it. By providing comprehensive spatial information through a singular model, we can eliminate the need to use multiple models to understand various aspects of the anomaly. Furthermore, for such a model to perform robustly, we must be able to train it on a diverse dataset. In the real world, different variants of the same disease can arise in different areas, thus requiring distributed training methods like Federated Learning (FL) to produce globally performant models.

In this chapter, we discuss the architecture of our proposed model, DiagnoFormer, which generates multiple diagnostic outputs for a single input image, enabling multifaceted anomaly localization. Using a pipelined Multi-Task Learning (MTL) architecture, our model integrates three key medical imaging tasks within a unified framework. Such exploitation of shared features and the usage of a hybrid loss function enables synergistic training of distinct tasks. Our approach minimizes task-specific data dependency and eliminates the need to train different models for specific diagnostic outputs. Additionally, the hybrid loss function ensures balanced training across tasks by preventing any single task from being prioritized. In this chapter, we also discuss the various FL techniques evaluated in our study, including our proposed Federated Bayesian Learning (FBE) method.

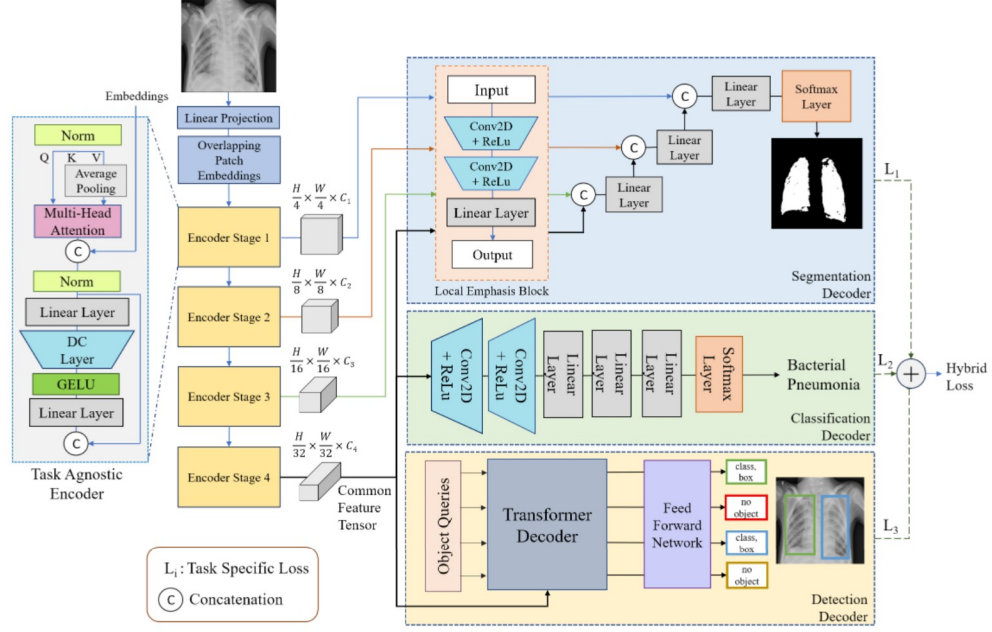


Figure 3.1: Architecture of DiagnoFormer. The encoder comprises multi-headed self-attention layers followed by a depthwise convolution layer (DC). The Encoder Stage 4 output is sent to all the task-specific decoders for generating respective outputs. The hybrid loss is calculated by taking a weighted sum of the task-specific losses from the decoders. (H, W, C) refer to height, width, and number of channels respectively.

Section 3.1 provides a detailed breakdown of the various components of DiagnoFormer. Section 3.2 discusses the various task-specific losses and the hybrid loss function used to train the model. Finally, Section 3.3 outlines the various federation methods studied and the mechanism of our proposed FBE method for multi-task imaging models.

3.1 Proposed Model: DiagnoFormer

DiagnoFormer employs a task-agnostic transformer encoder combined with task-specific decoders to generate multiple outputs. Our model features a custom transformer encoder designed to learn both local and global features in a task-agnostic manner. Additionally, we leverage Multi-Task Learning (MTL) and a hybrid loss function to enhance feature sharing, ensuring that the task-specific decoders are trained collaboratively. Figure 3.1 provides a detailed overview of our model’s architecture,

while Algorithm 1 presents the training procedure.

Algorithm 1 Our proposed model: DiagnoFormer.

Require: Input (X, Y) and Model $M(\text{Encoder } E, \text{Decoders } B_1, B_2, B_3)$.

- 1: **for** $b \in 1$ *to* n **do** \triangleright loop for each batch
- 2: $s = E(x_b \in X)$ $\triangleright s$: Output(s) of Encoder
- 3: **for** $i \in 1$ *to* 3 **do in parallel**
- 4: $\hat{p}_i = B_i(s_i \subseteq s)$ \triangleright compute task-specific predictions.
- 5: $l_i = f_i(\hat{p}_i, p_b \in Y)$ \triangleright compute task loss.
- 6: **end for**
- 7: $l_b^{hybrid} = \sum_{i=1}^3 (\sigma_i \cdot l_i)$ \triangleright compute hybrid loss.
- 8: $M(w^{t+1}) = \text{Backprop}(M(w^t), l_b^{hybrid})$ \triangleright update model weights.
- 9: **end for**

3.1.1 Task-Agnostic Encoder

The standard Vision Transformer encoder consists of multi-head attention (MHA) layers, followed by two fully connected Multi-Layer Perceptron (MLP) layers. In medical imaging, capturing local features is crucial for identifying key characteristics within the images. Thus, as per Zhang et al. (2021), we have incorporated a 3×3 depth-wise convolution (DC) layer followed by a GELU activation function, placing it between the two MLP layers. Assuming the output of the self-attention mechanism as SA , the convolution operation is performed as

$$O[p, q] = f \left((b[p, q] + \sum_{m=-\alpha}^{m=\alpha} \sum_{n=-\beta}^{n=\beta} SA[p+m][q+n] \cdot K[m, n]) \right) \quad (3.1)$$

where $f(\cdot)$ is the GELU activation function, K denotes the learnable convolution kernel with dimensions (m, n) , O is the output tensor with extracted feature channels, (p, q) are the pixel coordinates of any given pixel, and α, β are the receptive fields in the width and height dimensions, respectively.

The convolution layer captures local features that complement the global dependencies learned by the attention layers. This enables DiagnoFormer to handle tasks that require proper contextualization of micro and macro features, such as object detection. The feature tensor from the final layer of the encoder block, which retains

the highest number of feature channels, is passed to the three task-specific decoders.

3.1.2 Task-Specific Decoders

For segmentation, we have adopted the decoder architecture from SSFormer (Wang et al., 2022). The feature tensor produced by the last layer of the task-agnostic encoder is input into the Progressive Locality Decoder (PLD). The PLD consists of CNN layers and feature recombination operations as highlighted in Figure 3.1. The CNN layers upscale the feature tensors extracted at various encoder levels. The stepwise recombination operations combine the upscaled tensors to progressively highlight segmented patches. The final softmax layer produces the binary segmentation mask.

Object detection performance depends equally on the quality of both local and global features (Zhang et al., 2021). The task-agnostic encoder serves as a powerful feature extractor for object detection applications. In this regard, we have adopted the decoder architecture from DETR (Carion et al., 2020). Similar to segmentation, the encoder output is provided as input to the detection decoder. Additionally, a set of randomly initialized object queries is also provided as input. These object queries I are of the format $[I_{cx}, I_{cy}, I_h, I_w, I_{c_1 \dots c_n}, I_{conf}]$, where (I_{cx}, I_{cy}) are the coordinates of the centre point and (I_h, I_w) refer to the height and width of the bounding box respectively. $I_{c_1 \dots c_n}$ refers to which class the detected object belongs to, and finally, I_{conf} refers to the confidence score. The weights of these query tensors are learnable parameters that capture the position and dimension information of the objects detected within an image.

For image classification, we have implemented a custom architecture that includes two CNN layers, followed by three MLP layers, and a softmax layer. The CNN layers downsample the incoming feature tensor by condensing the number of feature channels. Subsequently, the MLP layers forward the condensed input through the network. The final output from the softmax layer is a single class label that predicts the type of disease.

3.2 Loss Functions

3.2.1 Task Specific Losses

For segmentation, we have employed Combo Loss (Zhang et al., 2021), which is a weighted sum of Dice Loss and Cross Entropy Loss. The Dice Coefficient addresses the class imbalance between pixels belonging to the foreground and background. Cross Entropy (CE) manages the trade-off between False Positives and False Negatives in pixel classification. For detection, we combine gIoU, L1, and CE losses to estimate one aggregate loss value as suggested by Carion et al. (2020). The gIoU and L1 losses assess the precision of the bounding box’s location predictions. CE is used to evaluate the accuracy of the object classifications within the predicted bounding boxes. For image classification, we utilize CE to determine the classification error across the various classes.

3.2.2 Hybrid Loss

For DiagnoFormer, we have utilized a combination of the aforementioned losses to calculate a single loss value for the entire model. Assuming task-specific losses $l_i \in \{l_1, l_2, l_3\}$ where l_1 : Segmentation Loss, l_2 : Detection Loss and l_3 : Classification Loss, and $\sigma_i \in \{\sigma_1, \sigma_2, \sigma_3\}$ as their respective weights, then

$$Hybrid\ Loss = \sum_{i=1}^3 \sigma_i \times l_i \quad (3.2)$$

We present the exact estimation of the hyperparameter values σ_i in section 4.4

3.3 Federation Schemes

In our research, we investigate two different yet popular approaches to federation: FedAvg (McMahan et al., 2017), and FedBE (Chen and Chao, 2021). We also evaluate a variation of FedAvg called Partial Participation Federation (PPF). PPF mimics a federation setting that resembles real-world conditions, wherein users only perform a

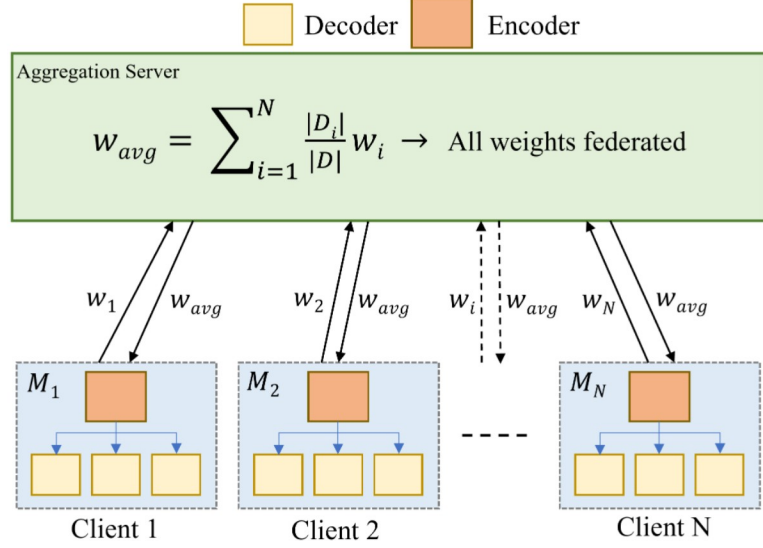


Figure 3.2: Overview of FedAvg. Given N clients, the server performs federated averaging and sends back w_{avg} to clients. Here M_i and w_i refer to client models and their respective locally trained weights. $|D_i|$ denotes the dataset size of client i and $|D| = \sum |D_i|$

subset of tasks.

3.3.1 Federated Averaging (FedAvg)

FedAvg (McMahan et al., 2017), is the most commonly implemented technique for model aggregation. In this method, a central server acts as the aggregator, averaging the model weights from various clients, weighted by the dataset size of each client. After aggregation, the averaged weights are distributed back to all clients for further local training or finetuning. This cycle repeats until the server model converges. An overview of the FedAvg process is depicted in Figure 3.2.

3.3.2 Partial Participation Federation (PPF)

A practical use-case in federated learning involves clients only performing a subset of tasks (Kim et al., 2022), a concept we refer to as Partial Participation. In this model, the encoder weights, denoted as w_i^{enc} , are federated separately. The decoder weights of those clients involved in a specific task are federated with other clients performing

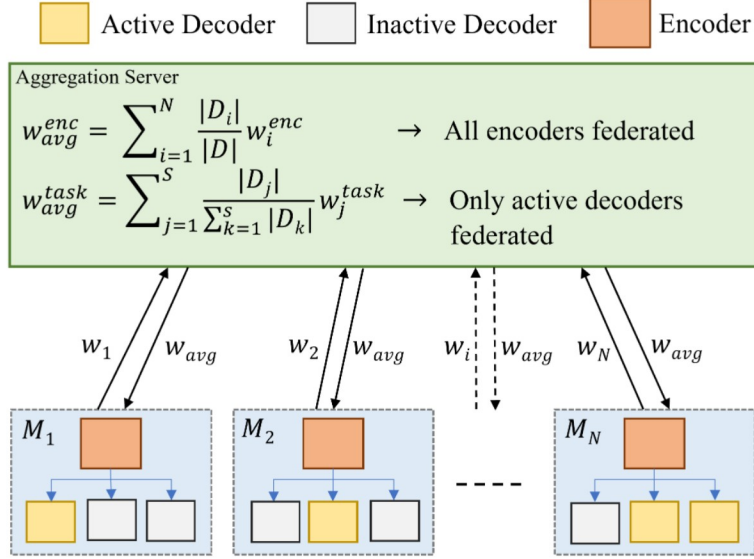


Figure 3.3: Overview of Partial Participation Federation. Encoder weights w_{avg}^{enc} and decoder weights w_{avg}^{task} are computed separately. M_i and w_i refer to client models and their respective locally trained weights, whereas D refers to the client dataset.

the same task. These task-specific weights are represented as w_j^{task} . Such selective federation targets efficiency by aligning the aggregation process with the specific tasks clients are engaged in. This process ensures that if a client does not perform a specific task, the parameters associated with their dormant decoder(s) are not considered for federation. Mathematically, the federated weights for the encoder and decoder are computed as:

$$w_{avg}^{enc} = \sum_{i=1}^N \frac{|D_i|}{|D|} w_{enc}^i \quad (3.3)$$

$$w_{task}^{avg} = \sum_{j=1}^S \frac{|D_j|}{\sum_{k=1}^S |D_k|} w_j^{task} \quad (3.4)$$

where N is the total number of clients and $S \subseteq N$ is the subset of clients participating in any given federation round for a specific task. $|D_i|$ signifies the dataset size of client i . w_{avg}^{enc} and w_{avg}^{task} refer to the computed federated encoder weights and federated task-specific decoder weights, respectively. An overview of the PPF aggregation process is given in Figure 3.3

3.3.3 Post-federation Finetuning

FL typically requires many rounds of federation to achieve collective convergence of global and local models. However, client convergence can be accelerated through client-side finetuning (Guan et al., 2024). Following the final round of federation, clients may perform a few extra training iterations locally. This allows the client models to personalize the shared federated knowledge by adjusting the global weights to better reflect their individual data distributions.

3.3.4 Federated Bayesian Ensemble (FBE)

As previously discussed, FedAvg encounters certain performance limitations. FedAvg is often ineffective in producing a performant global model, especially if the number of federation epochs is limited. While client-side finetuning may improve the performance of individual client models, the global model remains underperformant. Furthermore, FedAvg has been observed to deviate from the theoretical optimal model and may experience performance degradation due to the permutation-invariant nature of neural networks (Karimireddy et al., 2020). To address these issues and accelerate convergence, we propose a Bayesian federated learning approach called FedBE (Chen and Chao, 2021). FedBE circumvents the issues arising from direct averaging of weights by fitting a joint Gaussian distribution to the client weights. Subsequently, new weights are sampled from this distribution and knowledge distillation (Hinton et al., 2015) is used to train a singular robust global model. In our work, we have extended FedBE to perform segmentation alongside classification by modifying the classification algorithm to predict class probabilities on a per-pixel basis.

Algorithm 2 details the working of our proposed Bayesian Federation process. First, client model weights w_c are uploaded to the aggregation server after a few local training epochs. Then, T teacher models, each with associated weights $w^{(t)}$, where $t \in T$, are sampled from a unified Gaussian distribution fitted to client model weights w_c . Each teacher model t generates a segmentation mask $mask^{(t)}$ with dimensions $H \times W$ for an unlabeled server dataset U . To aggregate these feature maps into a final

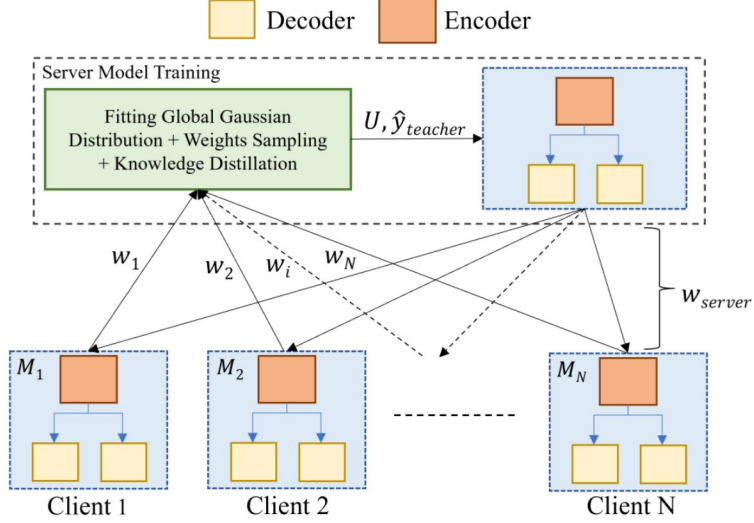


Figure 3.4: Overview of Federated Bayesian Ensembling. The client models upload their weights to the server where a Gaussian distribution is fitted to them. Teacher models are sampled from this distribution and used to train the server model through knowledge distillation. M_i are the client models, U refer to unlabeled data samples, and $\hat{y}_{teacher}$ refer to the pseudo-labels output by the sampled teacher models.

prediction, each pixel in a teacher's segmentation mask $mask_{(p,q)}^{(t)}$ is compared with the corresponding pixel from every other teacher's mask $mask_{(p,q)}^{(T-t)}$, where $p \in H$ and $q \in W$. Each teacher model provides a probabilistic prediction for the pixel's class: foreground or background. Finally, a mode function is applied to determine the most frequently predicted class for each pixel across all teacher models. The function is defined as:

$$mask_{(p,q)} = mode(mask_{(p,q)}^{(t)})_{t=1}^T \quad (3.5)$$

The final label for each pixel $mask_{(p,q)}$ is thus assigned based on a majority vote. This procedure is applied to every pixel across the segmentation maps generated by all teacher models. A similar procedure is followed on a per-prediction level to determine the class label $label$ for the image. The resulting unified segmentation mask $mask$ alongwith the classification label $label$ form the pseudo label $\hat{y}_{teacher} = (mask, label)$ that serves as the pseudo-ground truth for training the server model on the unlabeled dataset U . Figure 3.4 provides an overview of the Bayesian federation process. Due

to implementation challenges, we do not incorporate object detection in Bayesian Federation.

Algorithm 2 Federated Bayesian Ensemble (FBE) for Multi-Task Imaging.

Require: Server Input: model M , client weights w_c , unlabeled data U , labeled data D

- 1: **for** $r \in 1$ to R **do** \triangleright loop for each federation round
- 2: $w_g, w_f = \text{FedBE}(w_c, D)$ \triangleright get global weights w_g and FedAvg weights w_f using FedBE algorithm.
- 3: **Sample** T global models $\{w^{(t)} \sim p(w|D)\}_{t=1}^T$
- 4: **for teacher** $t \in 1$ to T **do**
- 5: $mask^{(t)}, label^{(t)} = M(w^{(t)}, U)$ \triangleright compute mask and label prediction for each sampled model $M(w^{(t)})$
- 6: **end for**
- 7: **for each pixel** $(p, q) \in \text{Image}(H, W)$ **do**
- 8: $mask_{(p,q)} = \text{mode}(mask_{(p,q)}^{(t)})_{t=1}^T$ \triangleright as per Equation 3.5
- 9: $mask_{(p,q)} = \text{softmax}(mask_{(p,q)})$
- 10: **end for**
- 11: $label = \text{mode}(label^{(t)})_{t=1}^T$ \triangleright compute majority prediction class label
- 12: $label = \text{softmax}(label)$
- 13: $\hat{y}_{teacher} = (mask, label)$
- 14: **Server Model Update:** $w = \text{SWA}(w_f, U, \hat{y}_{teacher})$ \triangleright train server model using Stochastic Weight Averaging (SWA) algorithm as per FedBE
- 15: **end for**

Chapter 4

Experimental Evaluation

In this chapter, we present the datasets, the architecture and the performance analysis of our proposed architecture, DiagnoFormer for medical imaging.

4.1 Dataset and Metrics

To compare our approach with existing multi-task models, we utilize the INBreast dataset (Moreira et al., 2012). INBreast is a full-field mammography dataset annotated by medical experts. Each image of a tumorous mass in the dataset has been assigned a BI-RADS (Breast Imaging Reporting and Data System) score (Eberl et al., 2006), which indicates whether it is malignant or benign. In line with comparable studies (Gao et al., 2020), we selected 108 annotated images with tumorous masses from the available 115 images for our evaluation.

For evaluating federated learning performance and benchmarking against single-task models, we use the v7-Labs Lung Segmentation dataset¹, which includes 6,500 high-resolution chest X-ray images. According to the official v7-Labs GitHub page², the images have been manually labeled by a human workforce from CloudFactory and reviewed by the v7 Labs team. Since our focus is on pneumonia and related conditions, we excluded 517 images associated with COVID-19 cases. Additionally, we removed

¹Dataset available at: <https://www.v7labs.com/open-datasets/covid-19-chest-x-ray-dataset>

²v7-Labs GitHub: <https://github.com/v7labs/covid-19-xray-dataset>

296 images due to incomplete or missing labels. The final dataset consists of 5,687 images with verified segmentation masks and classification labels for pneumonia.

For both datasets, bounding boxes for object detection were generated using OpenCV by drawing rectangles around the outer boundaries of the segmentation masks. We have followed an 80:20 split for training and testing sets across all tasks.

To evaluate model performance for each task, we use the following metrics:

- For segmentation, we have used Dice score (Jadon, 2020). Dice estimates the area of overlap between the predicted segmented patch and the ground truth. A score of 0 signifies no overlap whereas 1 means perfect segmentation.
- For detection, we utilize two metrics. Mean Average Precision (mAP) (Henderson and Ferrari, 2017) score with confidence thresholds ranging from 0.4 to 0.75 has been used for all experiments involving the v7-Labs dataset³. The average precision (AP) with a confidence threshold th is the area under the precision-recall curve for all predicted bounding boxes and their corresponding ground truth (GT) boxes. If a predicted box has an Intersection-over-Union (IoU) value $\geq th$ with a GT box, then the bounding box is considered a valid prediction. This AP value is computed for all values of $th \in \{0.4, 0.75\}$ with a step of 0.05. The average of all the AP values is reported as the final mAP score.

For comparison with models using the INBreast dataset, True Positive Rate @ False Positives per Image (TPR@FPI) has been used, following (Gao et al., 2020, Liu et al., 2024). This metric estimates the number of correctly predicted bounding boxes compared to the number of false positives predicted per image. The confidence threshold for a bounding box to be considered a prediction was set to 0.9.

- For classification, we measure performance using the Area Under Curve (AUC) (Huang and Ling, 2005) score. The AUC score measures the extent of separation between correct and incorrect predictions by taking the area under the True

³Kaggle Challenge: <https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/>

Table 4.1: Dataset division of v7-Labs dataset for all federated training and evaluation. PNA stands for Pneumonia. Kindly note that only training images are augmented.

Client	Purpose	Class			Total
		No PNA	Bacterial PNA	Viral PNA	
1	Train	0	0	2772	2772
	Test	0	0	178	178
2	Train	0	2851	140	2991
	Test	0	182	10	192
3	Train	0	3000	0	3000
	Test	0	192	0	192
4	Train	2976	0	0	2976
	Test	192	0	0	192
5	Train	777	785	499	2061
	Test	51	50	32	133
Total	Train	3753	6636	3411	12800
	Test	243	424	220	887

Positive - False Positive curve derived from the confusion matrix. The higher the separation, the better the AUC score.

4.2 Organization of Federation

For our federated learning experiments, we have assumed that each client retains their ground truth images, segmentation masks, and classification labels in a private and secure manner. Before commencement of training, an initial set of model weights is distributed from the server to each client model as per McMahan et al. (2017).

We have used the v7-Labs dataset for all FL experiments and assumed 5 participating clients. The images are distributed among these clients in a non-i.i.d (non-independent and identically distributed) fashion. Specifically, each client is assumed to have a disproportionate number of images pertaining to one type of pneumonia over others. This distribution strategy mirrors a real-world scenario where certain geographical regions may see the prevalence of one specific disease variant. Table 4.1 details the exact split of images among clients.

Table 4.2: Augmentations applied to training dataset. H. Flip is Horizontal Flip. \checkmark/\times represents instances where the augmentation has been randomly applied half the time.

Augmentation	Rotation	Shear	H. Flip	Gaussian Noise
Scheme 1	$\leq 5.0^\circ$	$\leq 2.5^\circ$	\times	\times
Scheme 2	$\leq 6.0^\circ$	$\leq 4.0^\circ$	\checkmark/\times	\times
Scheme 3	$\leq 6.0^\circ$	$\leq 4.0^\circ$	\checkmark/\times	\checkmark/\times

4.3 Dataset Augmentation

To enhance the quantity and diversity of training samples, we have employed a few mild image augmentations. These include a combination of rotation, shearing, horizontal flipping, and Gaussian noise addition. We have structured the augmentations into three distinct schemes. The details of the schemes with their respective parameters have been given in Table 4.2. For rotation and shearing, random values were generated within the specified maximum range. Horizontal flipping and Gaussian noise were each applied to approximately 50% of the images, contingent upon the specific augmentation scheme. To maintain the integrity of the dataset, we discarded any image in which key anatomical features were significantly distorted or truncated as a result of augmentation.

The INBreast dataset has an approximate benign-to-malignant tumour ratio of 1:3. To address this class imbalance, we applied a second round of augmentations to the benign tumour images generated during the initial augmentation pass. For secondary augmentations, the extent of rotation and shearing was limited to no more than $1.5\times$ the original maximum parameter values. Kindly note that no additional Gaussian noise was introduced during this phase. The final INBreast training set comprises 2080 malignant and 1840 benign tumour images. For the v7-Labs dataset, a single round of all augmentation strategies has been applied to each image, resulting in a final training set of 12,800 augmented images.

4.4 Choice of Training Parameters

For all experiments, DiagnoFormer has been trained with the AdamW (Loshchilov and Hutter, 2018) optimizer and a learning rate of 1×10^{-4} .

In estimating the hyperparameters for our hybrid loss function, experimentally, we noted that the segmentation and classification losses typically fluctuate between 0 and 2.5. In contrast, the detection loss often surpassed 7 in the early training epochs. This discrepancy is problematic for overall model performance, as it biases the model towards the detection task, leading to disproportionately large gradient steps. In some instances, this imbalance caused the segmentation and classification tasks to diverge.

To address this issue and ensure that all tasks are weighted equitably, we conducted grid search to optimize the scaling factors for each task-specific loss. Through this process, we identified the optimal set of scaling factors, or σ values, to be $[1, 0.375, 1]$ for segmentation, detection, and classification, respectively. These adjustments help normalize the influence of the task-specific losses, promoting balanced learning across all tasks.

4.5 Comparative Evaluation

In this section, we evaluate the performance of DiagnoFormer against single-task and multi-task models. We also perform a detailed analysis of DiagnoFormer under various FL scenarios. Finally we show a parameter count comparison against state-of-the-art multi-task transformer models.

4.5.1 Performance Analysis of DiagnoFormer

To demonstrate DiagnoFormer’s effectiveness in pneumonia detection, we compare its results against widely used single-task models: SSFormer (Wang et al., 2022) for segmentation, DETR (Carion et al., 2020) for detection, and ViT (Dosovitskiy et al., 2020) for classification. Additionally, to assess detection and multi-task performance on the INBreast dataset, we benchmark our model against GF-FPN (Liu et al.,

Table 4.3: Performance of DiagnoFormer against existing models.

Model	Dataset	Dice	mAP	TPR	AUC
SSFormer (Wang et al., 2022)	v7-Labs	0.952	X	-	X
DETR (Carion et al., 2020)	v7-Labs	X	0.834	-	X
ViT (Dosovitskiy et al., 2020)	v7-Labs	X	X	-	0.850
OURS	v7-Labs	0.952	0.882	-	0.930
GF-FPN (Liu et al., 2024)	INBreast	X	-	0.96 (FPI=0.56)	X
ResCU-Net (Shen et al., 2019)	INBreast	0.917	-	X	0.961
FT-MTL-Net (Gao et al., 2020)	INBreast	0.760	-	0.91 (FPI=1.5)	0.920
OURS	INBreast	0.792	-	0.96 (FPI=0.54)	0.962

Note: Dice Score is used for segmentation, mAP (0.4-0.75) and TPR@FPI for detection, and AUC for classification. - represents cases where a metric is not applicable, whereas **X** is used where a model does not perform a specific task. In each case, higher is better.

2024), FT-MTL-Net (Gao et al., 2020), and ResCU-Net (Shen et al., 2019). Table 4.3 presents the quantitative results of our experiments, showing that DiagnoFormer performs competitively against both single-task and multi-task models across all three tasks. Notably, DiagnoFormer consistently delivers superior results in detection and classification on the INBreast dataset.

DiagnoFormer’s multi-task performance can be attributed to the transformer encoder’s ability to learn robust task-agnostic features that generalize effectively across multiple tasks. As illustrated in Figure 4.1, DiagnoFormer accurately segments small, scattered cancerous tumour patches within complex mammography images. Its ability to efficiently contextualize local features enables the model to disregard non-tumourous contours in the INBreast dataset correctly. With a limited INBreast dataset, DiagnoFormer achieves a Dice Score of 0.792, outperforming its closest three-task counterpart, FT-MTL-Net.

FT-MTL-Net maintains stable performance for object detection at an FPI of 1.5, while GF-FPN achieves stability at $FPI = 0.56$. In contrast, DiagnoFormer demonstrates superior and more stable detection performance with $TPR @ FPI = 0.96 @ 0.54$. This demonstrates that DiagnoFormer produces the fewest false positives per image among all competing models while achieving a true positive detection score of 0.96. This score is at par with GF-FPN, which is the current state-of-the-art model for

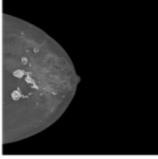


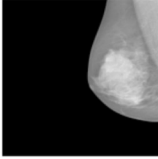


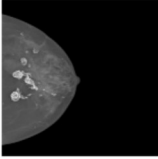
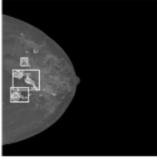
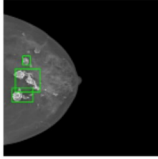
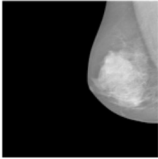
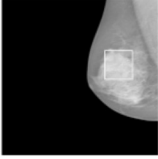
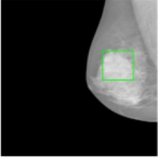
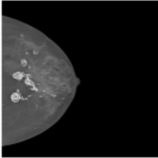
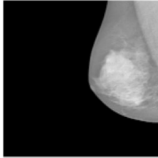
<div>Segmentation</div> <div>Detection</div> <div>Classification</div>	Input Image	Ground Truth	Predicted	Input Image	Ground Truth	Predicted
						
						
		Benign Tumour	Benign Tumour		Malignant Tumour	Malignant Tumour

Figure 4.1: Qualitative results of DiagnoFormer on INBreast dataset

tumour detection on the INBreast dataset. Additionally, our DiagnoFormer attains the highest overall AUC score among all multi-task classification models.

DiagnoFormer also outperforms popular single-task models, matching the segmentation performance of its counterpart SSFormer while exceeding DETR in detection and ViT in classification. These performance improvements can be attributed to the synergistic training of task-specific decoders, allowing them to learn from each other. Figure 4.2 illustrates DiagnoFormer’s qualitative results on the v7-Labs Pneumonia dataset.

4.5.2 Analysis of Federation Methods

Table 4.4 showcases the efficacy of our averaging schemes compared to data-centralized learning using DiagnoFormer. Our findings underscore how effectively DiagnoFormer adapts to FL techniques, managing to surpass centralized learning performance for segmentation and classification. Additionally, the table highlights the significant impact of localized finetuning post federation.

For each evaluation metric, we report the average score from the best performing









Task	Segmentation	Input Image	Ground Truth	Predicted	Input Image	Ground Truth	Predicted
	Detection						
	Classification		Bacterial Pneumonia	Bacterial Pneumonia		No Pneumonia	No Pneumonia

Figure 4.2: Qualitative results of DiagnoFormer on v7-Labs Darwin dataset

finetuning epoch following the 5th round of federation across clients. In the following paragraphs, we present a detailed analysis of the tasks performed:

Table 4.4 highlights the effectiveness of our proposed Bayesian Ensembling method (FBE) for segmentation tasks. Without any local finetuning, FBE achieves the best Dice score of 0.841, significantly outperforming other federation methods. After finetuning, the Dice score improves to 0.953. Such strong performance is largely due to FBE’s ability to overcome the issue of permutation invariance during the creation of the global model. The results also underscore the value of local finetuning in enhancing task-specific accuracy.

Unlike segmentation, object detection suffers considerably when transitioning from centralized to federated training. As shown in Table 4.4, the mean Average Precision (mAP) score drops from 0.882 under centralized training to 0.396 under FedAvg. Although finetuning helps improve client performance, the gains are modest. This performance degradation stems from the averaging of permutation-invariant object queries. In the detection decoder, the object queries are randomly initialized, with each query targeting a different region of the image. Directly averaging these independently trained queries leads to distorted feature representations. Given that object detection depends heavily on the interplay between local and global features, finetuning is unable

Table 4.4: Performance of DiagnoFormer under data centralized, and various federated and finetuned settings.

Strategy	Finetuning	Dice	mAP	AUC
Data Centralized	-	0.952	0.882	0.930
Federated Averaging (FedAvg)	X	0.532	0.396	0.500
	✓	0.951	0.513	0.952
Partial Participation Federation (PPF)	X	0.523	0.511	0.603
	✓	0.950	0.618	0.784
Federated Bayesian Ensembling (FBE)	X	0.841	-	0.556
	✓	0.953	-	0.728

Note: Dice Score is used for segmentation, mAP (0.4-0.75) for detection, and AUC for classification. In each case, higher is better.

to recover performance adequately. In this context, PPF with finetuning yields the best results, achieving an average client mAP of 0.511. This score improves to 0.618 after finetuning. We surmise that PPF’s superior performance can be attributed to the reduced number of participating clients for the object detection task compared to the total client pool.

For image classification, FBE outperforms FedAvg in terms of AUC score. FBE achieves an AUC score of 0.556 as compared to 0.500 for FedAvg. However, it falls short of PPF presumably due to higher client participation in FBE. While local finetuning results in a higher AUC after FedAvg and PPF, the gains after FBE are also notable.

4.5.3 Model Sizes and Communication Benefits

To evaluate the communication efficiency of DiagnoFormer under federated settings, we compare its overall model size and communication overhead against FeSTA (Park et al., 2021): a multi-task transformer-based model performing equivalent tasks. Comparing DiagnoFormer’s performance against FeSTA is not feasible as the latter uses distinct inputs for each task. This differs from our use case, where the input is a singular image. However, comparing parameter count between the two provides valuable insights into their operational efficiency in a federated setting.

Table 4.5: Model Sizes and Parameter Count Comparison.

Model	Param Type	Task	Size (MB)	Total Size (MB)	Params (M)	Total Params (M)
FeSTA (Park et al., 2021)	Federated	Segment	89.8	331.9	22.428	148.968
		Detect	187.9		46.858	
		Classify	54.2		13.315	
	Server	Model	265.5	265.5	66.367	
OURS	Federated	Segment	20.5	159.8	5.376	41.906
		Detect	37.1		9.740	
		Classify	9.9		2.594	
		Encoder	92.3		24.196	

The analysis focuses on two key aspects: the total size of the model and the number of parameters that need to be communicated during the federation process. Our findings demonstrate that DiagnoFormer offers substantial advantages in both aspects. Table 4.5 clearly illustrates the differences in the parameter count and the disk space occupied by DiagnoFormer compared to FeSTA. The data shows that the total number of parameters DiagnoFormer communicates during a federation round is nearly half that of FeSTA. This reduction in communication load is highly beneficial for FL, where bandwidth is often a crucial limitation.

4.6 Ablation Studies

4.6.1 Replacing Transformer Encoder with CNN UNet

In this section, we examine the impact of the transformer backbone on the overall performance of our model. For this experiment, we replace the encoder and segmentation decoder with an equivalent popular CNN-based architecture, UNet (Ronneberger et al., 2015). This replacement allows us to directly compare the effectiveness of a transformer-based approach against a CNN-based approach within the same framework.

Table 4.6: Performance variation between clients using the CNN-based model vs DiagnoFormer.

Model →		CNN-Encoder Multi-Task						DiagnoFormer					
Parameter Type →		Centralized			Federated			Centralized			Federated		
Client ↓		Dice	mAP	AUC	Dice	mAP	AUC	Dice	mAP	AUC	Dice	mAP	AUC
1		0.216	0.126	0.607	0.867	0.151	0.826	0.948	0.865	0.880	0.949	0.491	0.900
2		0.274	0.055	0.806	0.831	0.072	0.762	0.951	0.863	0.815	0.951	0.515	0.912
3		0.453	0.135	0.705	0.902	0.045	0.811	0.952	0.893	0.884	0.952	0.493	0.895
4		0.542	0.264	0.805	0.883	0.241	0.767	0.948	0.901	0.888	0.949	0.531	0.916
5		0.857	0.067	0.713	0.762	0.039	0.905	0.951	0.876	0.885	0.951	0.616	0.911

Note: Federation is applied for only 1 communication round. Dice Score is used for segmentation, mAP (0.4-0.75) for detection, and AUC for classification. In each case, higher is better.

4.6.1.1 Performance analysis of the two models

For this experiment, we have utilized a common test set of 887 images from the v7-Labs dataset, distributed across three classes: 243 for No Pneumonia, 424 for Bacterial Pneumonia, and 220 for Viral Pneumonia. The training data distribution among clients has been kept consistent with previously discussed FL experiments.

Each client model has been trained on their local dataset and scores have been reported after 200 training epochs. For non-federated experiments, as presented in Table 4.6, the CNN-based model achieves an average Dice Score of 0.468 and an mAP of 0.129. In contrast, DiagnoFormer significantly outperforms it with scores of 0.95 and 0.879, respectively. This disparity arises from the visual differences between the positive and negative pneumonia samples. These differences are evident on a macro level, which the CNN model struggles to contextualize due to its limited ability to capture global features. Hence, the CNN-based model exhibits considerable variability in performance across individual clients and overall poorer performance. However, for classification, the CNN-based model’s performance is comparable to that of DiagnoFormer. This is because image classification is a relatively straightforward task that does not heavily rely on high-quality features. Nevertheless, DiagnoFormer’s superior feature learning ability enables it to maintain better performance parity across clients. The difference in AUC scores between the best and worst-performing clients is 0.199 for the CNN-based model, whereas for DiagnoFormer, it is 0.073.

4.6.1.2 Effect of one-shot federation on model performance

In this section, we examine the impact of one-shot federation on mitigating the effects of non-i.i.d. data distribution among clients. FedAvg is used to aggregate the weights of all clients following 200 epochs of local training. As detailed in Table 4.6, applying federation significantly enhances the performance of our CNN-based model. We observe a significant increase in accuracy and a decrease in performance difference between the best and worst-performing clients.

Following just one round of federation and subsequent finetuning, the average client Dice score improves to 0.848, while the score difference between the best and worst-performing clients reduces from 0.641 to 0.140. Similarly, classification performance also improves, with the average AUC score reaching 0.814. These improvements highlight the benefit of knowledge sharing through parameter averaging during federation. Parameter averaging allows each client model to gain insights into the data distributions of other clients, leading to the learning of robust features. However, object detection does not benefit accordingly, exhibiting a behaviour consistent with prior FL experiments.

In contrast, DiagnoFormer shows more modest improvements in accuracy and performance difference between the best and worst-performing clients for segmentation and classification. For DiagnoFormer, FL predominantly refines the agnosticity of learned features rather than significantly improving baseline performance.

4.6.2 Analysis of Differential Privacy

While FL enhances privacy by limiting direct data sharing, the periodic transmission of gradients exposes models to potential security threats, including actively malicious attacks (Fowl et al., 2022, Lu et al., 2022) and gradient inversion tactics (Geiping et al., 2020). Differential Privacy (DP) (Dwork, 2006) mathematically limits how much information an adversary can extract from a model’s output. This protection is quantified by the parameter ϵ , with privacy safeguarded by injecting meticulously calibrated noise during training. This helps in striking a balance between

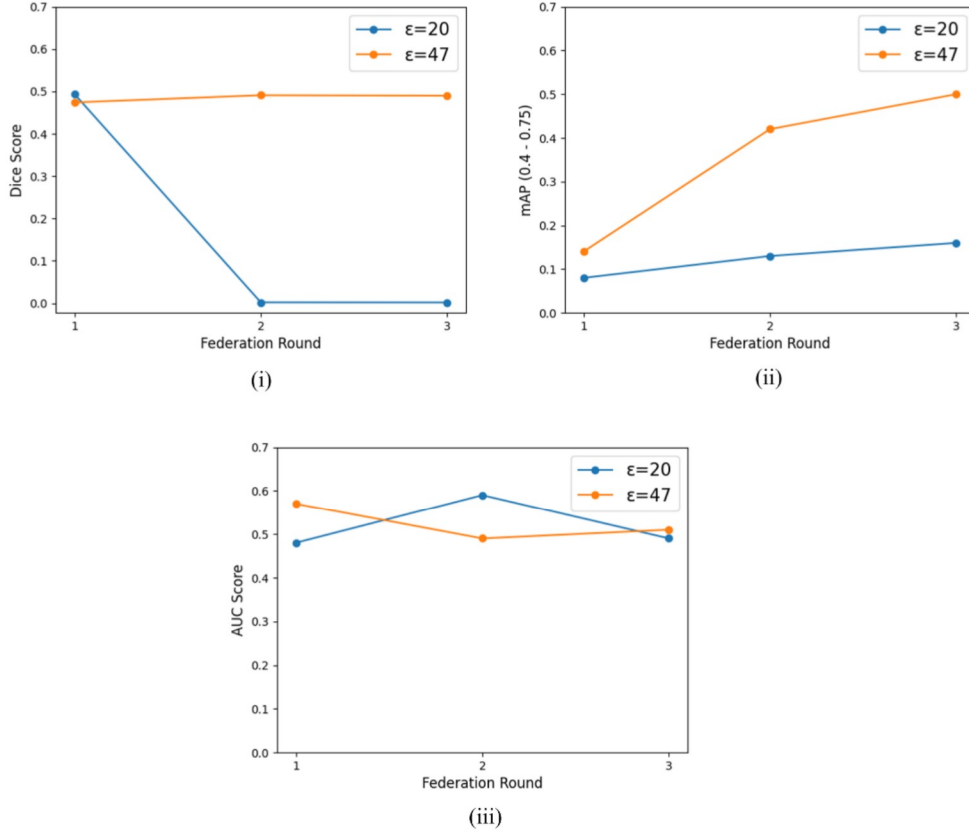


Figure 4.3: Overview of DP-SGD with $\epsilon = 20$ and $\epsilon = 47$. X-Axis: Federation Round, Y-Axis: Metric (i) Segmentation (Dice Score) (ii) Detection (mAP) (iii) Classification (AUC).

privacy and accuracy. Kindly note that a higher ϵ value indicates a lower privacy guarantee.

In practice, DP is implemented during model training using Differentially Private Stochastic Gradient Descent (DP-SGD) (Abadi et al., 2016). Upon training with DP-SGD, we observe that for a model like ours, where feature maps are of utmost importance, ensuring privacy comes at the cost of accuracy.

In our experiments, we average the accuracies of five clients over three federation rounds to examine how different ϵ values impact performance. Figure 4.3 illustrates the performance of DiagnoFormer at various ϵ values for each task. For segmentation, at $\epsilon = 20$, the model diverges after one federation round. Upon increasing ϵ to

47, segmentation performance stabilizes at a Dice score of 0.5, though no significant improvements are observed with further rounds. Higher ϵ values compromise privacy without yielding significantly better results. For object detection, $\epsilon = 47$ results in mAP scores up to 0.51. However, further training leads to inconsistent outcomes due to DP noise being added during each training batch.

Classification performance is less affected by variations in DP noise owing to the simplicity of the classification decoder. For both ϵ values of 20 and 47, classification performance remains consistent around an AUC score of 0.5, indicating minimal impact from DP adjustments but no significant performance improvement with a higher ϵ value.

Chapter 5

Conclusion and Future Scope

In this thesis, we have presented DiagnoFormer, a multi-task transformer model designed for simultaneous image segmentation, object detection, and classification. We have demonstrated that DiagnoFormer performs competitively across these tasks while having a lightweight architecture. We have also explored the model’s performance with federated learning by experimenting with various federation schemes. Furthermore, we have introduced a Bayesian Learning approach to semantic segmentation and classification for multi-task imaging models and illustrated its superiority in maximizing the Dice score compared to other methods.

We would like to highlight some limitations of our approach. Despite their impressive capabilities, vision transformers are resource-heavy models. This poses challenges for deployment on resource-constrained medical devices. Nonetheless, with the ongoing advancements in the processing power of handheld and wearable devices, we are optimistic that these hurdles will be mitigated in the near future. Additionally, successful large-scale implementation of federated learning requires a consistent, reliable, and secure internet connection between the client models and the aggregation server. In a real-world setting, this may be a pain point for clients situated in remote or underserved regions.

In the future, we aim to delve deeper into the privacy issues associated with models such as DiagnoFormer. In addition to the aforementioned issues, we would like to assess the efficacy of adversarial attacks in breaching patient data security in a federated

setting. We believe that addressing such issues is crucial for distributed training frameworks such as federated learning, where model parameters are frequently exchanged. Additionally, we plan to explore how DiagnoFormer performs when deployed across clients with varying datasets.

Bibliography

- Abadi M, Chu A, Goodfellow I, McMahan HB, Mironov I, Talwar K, Zhang L (2016) Deep Learning with Differential Privacy. In: Proc. ACM SIGSAC Conference on Computer and Communications Security, pp 308–318, DOI (10.1145/2976749.2978318), URL <https://doi.org/10.1145/2976749.2978318>
- Akselrod-Ballin A, Karlinsky L, Alpert S, Hasoul S, Ben-Ari R, Barkan E (2016) A Region Based Convolutional Network for Tumor Detection and Classification in Breast Mammography. In: Proc. Deep Learning and Data Labeling for Medical Applications, Springer, pp 197–205, DOI (10.1007/978-3-319-46976-8_21), URL https://doi.org/10.1007/978-3-319-46976-8_21
- Can YS, Ersoy C (2021) Privacy-preserving federated deep learning for wearable IoT-based biomedical monitoring. ACM Transactions on Internet Technology (TOIT) 21:1–17, DOI (10.1145/3428152), URL <https://doi.org/10.1145/3428152>
- Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S (2020) End-to-End Object Detection with Transformers. In: Proc. European Conference on Computer Vision (ECCV), Springer, pp 213–229, DOI (10.1007/978-3-030-58452-8_13), URL https://doi.org/10.1007/978-3-030-58452-8_13
- Chakraborty C, Khosravi MR, Casalino G, Rodrigues JJ (2023) Guest Editorial Special Issue on AIoMT-Enabled Federated Learning-Based Computing for Socially Implemented IoMT Systems: How Will Healthcare Systems Change? IEEE Transactions on Computational Social Systems 10:1537–1539, DOI (10.1109/TCSS.2023.3293352), URL <https://doi.org/10.1109/TCSS.2023.3293352>

- Chen H, Wang Y, Guo T, Xu C, Deng Y, Liu Z, Ma S, Xu C, Xu C, Gao W (2021) Pre-Trained Image Processing Transformer. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, pp 12299–12310, DOI (10.1109/CVPR46437.2021.01212), URL <https://doi.org/10.1109/CVPR46437.2021.01212>
- Chen HY, Chao WL (2021) FedBE: Making Bayesian Model Ensemble Applicable to Federated Learning. In: Proc. International Conference on Learning Representations (ICLR), OpenReview.net, URL <https://openreview.net/forum?id=dgtpE6gKjHn>
- Dayan I, Roth HR, Zhong A, Harouni A, Gentili A, Abidin AZ, Liu A, Costa AB, Wood BJ, Tsai CS, et al. (2021) Federated learning for predicting clinical outcomes in patients with COVID-19. *Nature Medicine* 27:1735–1743, DOI (10.1038/s41591-021-01506-3), URL <https://doi.org/10.1038/s41591-021-01506-3>
- Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, et al. (2020) An image is worth 16x16 words: Transformers for image recognition at scale. In: Proc. International Conference on Learning Representations (ICLR), OpenReview.net, URL <https://openreview.net/forum?id=YicbFdNTTy>
- Dwork C (2006) Differential privacy. In: Proc. International Colloquium on Automata, Languages, and Programming (ICALP), Springer, pp 1–12, DOI (10.1007/11787006_1), URL https://doi.org/10.1007/11787006_1
- Eberl MM, Fox CH, Edge SB, Carter CA, Mahoney MC (2006) BI-RADS classification for management of abnormal mammograms. *The Journal of the American Board of Family Medicine* 19:161–164, DOI (10.3122/jabfm.19.2.161), URL <https://doi.org/10.3122/jabfm.19.2.161>
- Filipiuk M, Singh V (2022) Comparing vision transformers and convolutional nets for safety critical systems. In: Proc. SafeAI@ AAI, URL https://ceur-ws.org/Vol-3087/paper_31.pdf

- Fowl LH, Geiping J, Czaja W, Goldblum M, Goldstein T (2022) Robbing the Fed: Directly Obtaining Private Data in Federated Learning with Modified Models. In: Proc. International Conference on Learning Representations (ICLR), OpenReview.net, URL <https://openreview.net/forum?id=fwzUgo0FM9v>
- Gao F, Yoon H, Wu T, Chu X (2020) A feature transfer enabled multi-task deep learning model on medical imaging. Expert Systems with Applications 143:112957, DOI (10.1016/j.eswa.2019.112957), URL <https://doi.org/10.1016/j.eswa.2019.112957>
- Geiping J, Bauermeister H, Dröge H, Moeller M (2020) Inverting Gradients - How easy is it to break privacy in federated learning? Advances in Neural Information Processing Systems (NeurIPS) 33:16937–16947, URL https://proceedings.neurips.cc/paper_files/paper/2020/file/c4ede56bbd98819ae6112b20ac6bf145-Paper.pdf
- Ghosh S, Ghosh SK (2023) FEEL: FEderated LEarning Framework for ELderly Healthcare Using Edge-IoMT. IEEE Transactions on Computational Social Systems 10:1800–1809, DOI (10.1109/TCSS.2022.3233300), URL <https://doi.org/10.1109/TCSS.2022.3233300>
- Graham S, Vu QD, Jahanifar M, Raza SEA, Minhas F, Snead D, Rajpoot N (2023) One model is all you need: Multi-task learning enables simultaneous histology image segmentation and classification. Medical Image Analysis 83:102685, DOI (10.1016/j.media.2022.102685), URL <https://doi.org/10.1016/j.media.2022.102685>
- Guan H, Yap PT, Bozoki A, Liu M (2024) Federated learning for medical image analysis: A survey. Pattern Recognition p 110424, DOI (10.1016/j.patcog.2024.110424), URL <https://doi.org/10.1016/j.patcog.2024.110424>
- Gurcan MN, Boucheron LE, Can A, Madabhushi A, Rajpoot NM, Yener B (2009) Histopathological Image Analysis: A Review. IEEE Reviews in Biomedical Engineering 2:147–171, DOI (10.1109/RBME.2009.2034865), URL <https://doi.org/10.1109/RBME.2009.2034865>

- Henderson P, Ferrari V (2017) End-to-End Training of Object Class Detectors for Mean Average Precision. In: Proc. Asian Conference on Computer Vision (ACCV), Springer, pp 198–213, DOI (10.1007/978-3-319-54193-8_13), URL https://doi.org/10.1007/978-3-319-54193-8_13
- Hinton G, Vinyals O, Dean J (2015) Distilling the knowledge in a neural network. In: Proc. NIPS Deep Learning and Representation Learning Workshop, DOI (10.48550/arXiv.1503.02531), URL <http://arxiv.org/abs/1503.02531>
- Huang J, Ling CX (2005) Using AUC and accuracy in evaluating learning algorithms. IEEE Transactions on Knowledge and Data Engineering 17:299–310, DOI (10.1109/TKDE.2005.50), URL <https://doi.org/10.1109/TKDE.2005.50>
- Jadon S (2020) A survey of loss functions for semantic segmentation. In: Proc. IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), IEEE, pp 1–7, DOI (10.1109/CIBCB48159.2020.9277638), URL <https://doi.org/10.1109/CIBCB48159.2020.9277638>
- Jochems A, Deist TM, El Naqa I, Kessler M, Mayo C, Reeves J, Jolly S, Matuszak M, Ten Haken R, van Soest J, et al. (2017) Developing and Validating a Survival Prediction Model for NSCLC Patients Through Distributed Learning Across 3 Countries. International Journal of Radiation Oncology* Biology* Physics 99:344–352, DOI (10.1016/j.ijrobp.2017.04.021), URL <https://doi.org/10.1016/j.ijrobp.2017.04.021>
- Karimireddy SP, Kale S, Mohri M, Reddi S, Stich S, Suresh AT (2020) Scaffold: Stochastic controlled averaging for federated learning. In: Proc. International Conference on Machine Learning (ICML), PMLR, pp 5132–5143, URL <http://proceedings.mlr.press/v119/karimireddy20a/karimireddy20a.pdf>
- Kim B, Kim J, Ye JC (2022) Task-Agnostic Vision Transformer for Distributed Learning of Image Processing. IEEE Transactions on Image Processing 32:203–218, DOI (10.1109/TIP.2022.3226892), URL <https://doi.org/10.1109/TIP.2022.3226892>

- Li F, Zhou L, Wang Y, Chen C, Yang S, Shan F, Liu L (2022) Modeling long-range dependencies for weakly supervised disease classification and localization on chest X-ray. *Quantitative Imaging in Medicine and Surgery* 12:3364–3378, DOI (10.21037/qims-21-1117), URL <https://doi.org/10.21037/qims-21-1117>
- Liu W, Zeng P, Jiang J, Chen J, Chen L, Hu C, Jian W, Diao X, Wang X (2024) Improved PAA algorithm for breast mass detection in mammograms. *Computer Methods and Programs in Biomedicine* 251:108211, DOI (10.1016/j.cmpb.2024.108211), URL <https://doi.org/10.1016/j.cmpb.2024.108211>
- Loshchilov I, Hutter F (2018) Decoupled weight decay regularization. In: *Proc. International Conference on Learning Representations (ICLR)*, OpenReview.net, URL <https://openreview.net/pdf?id=Bkg6RiCqY7>
- Lu J, Goswami V, Rohrbach M, Parikh D, Lee S (2020) 12-in-1: Multi-Task Vision and Language Representation Learning. In: *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE Computer Society, pp 10437–10446, DOI (10.1109/CVPR42600.2020.01045), URL <https://doi.org/10.1109/CVPR42600.2020.01045>
- Lu J, Zhang XS, Zhao T, He X, Cheng J (2022) APRIL: Finding the Achilles’ Heel on Privacy for Vision Transformers. In: *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE Computer Society, pp 10051–10060, DOI (10.1109/CVPR52688.2022.00981), URL <https://doi.org/10.1109/CVPR52688.2022.00981>
- McMahan B, Moore E, Ramage D, Hampson S, y Arcas BA (2017) Communication-efficient learning of deep networks from decentralized data. In: *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, PMLR, pp 1273–1282, URL <https://proceedings.mlr.press/v54/mcmahan17a/mcmahan17a.pdf>
- Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, Burren Y, Porz N, Slotboom J, Wiest R, et al. (2014) The Multimodal Brain Tumor

- Image Segmentation Benchmark (BRATS). IEEE Transactions on Medical Imaging 34:1993–2024, DOI (10.1109/TMI.2014.2377694), URL <https://doi.org/10.1109/TMI.2014.2377694>
- Moreira IC, Amaral I, Domingues I, Cardoso A, Cardoso MJ, Cardoso JS (2012) IN-breast: Toward a Full-field Digital Mammographic Database. Academic Radiology 19:236–248, DOI (10.1016/j.acra.2011.09.014), URL <https://doi.org/10.1016/j.acra.2011.09.014>
- Park S, Kim G, Kim J, Kim B, Ye JC (2021) Federated split task-agnostic vision transformer for COVID-19 CXR diagnosis. Advances in Neural Information Processing Systems (NeurIPS) 34:24617–24630, URL https://proceedings.neurips.cc/paper_files/paper/2021/file/ceb0595112db2513b9325a85761b7310-Paper.pdf
- Raparthi M, Dodda SB, Maruthi S (2021) AI-Enhanced Imaging Analytics for Precision Diagnostics in Cardiovascular Health. European Economic Letters (EEL) 11, DOI (10.52783/eel.v11i1.1084), URL <https://doi.org/10.52783/eel.v11i1.1084>
- Ronneberger O, Fischer P, Brox T (2015) U-Net: Convolutional Networks for Biomedical Image Segmentation . In: Proc. Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer, pp 234–241, DOI (10.1007/978-3-319-24574-4_28), URL https://doi.org/10.1007/978-3-319-24574-4_28
- Santos MK, Ferreira Júnior JR, Wada DT, Tenório APM, Nogueira-Barbosa MH, Marques PMdA (2019) Artificial intelligence, machine learning, computer-aided diagnosis, and radiomics: advances in imaging towards to precision medicine. Radiologia Brasileira 52:387–396, DOI (10.1590/0100-3984.2019.0049), URL <https://doi.org/10.1590/0100-3984.2019.0049>
- Sheller MJ, Edwards B, Reina GA, Martin J, Pati S, Kotrotsou A, Milchenko M, Xu W, Marcus D, Colen RR, et al. (2020) Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Scientific

reports 10:1–12, DOI (10.1038/s41598-020-69250-1), URL <https://doi.org/10.1038/s41598-020-69250-1>

Shen T, Gou C, Wang J, Wang FY (2019) Simultaneous Segmentation and Classification of Mass Region From Mammograms Using a Mixed-Supervision Guided Deep Model. *IEEE Signal Processing Letters* 27:196–200, DOI (10.1109/LSP.2019.2963151), URL <https://doi.org/10.1109/LSP.2019.2963151>

Wang J, Huang Q, Tang F, Meng J, Su J, Song S (2022) Stepwise Feature Fusion: Local Guides Global. In: *Proc. Medical Image Computing and Computer Assisted Intervention (MICCAI)*, Springer, pp 110–120, DOI (10.1007/978-3-031-16437-8_11), URL https://doi.org/10.1007/978-3-031-16437-8_11

Zhang D, Li J, Li X, Du Z, Xiong L, Ye M (2021) Local–Global Attentive Adaptation for Object Detection. *Engineering Applications of Artificial Intelligence* 100:104208, DOI (10.1016/j.engappai.2021.104208), URL <https://doi.org/10.1016/j.engappai.2021.104208>

Zhang M, Wang Y, Luo T (2020) Federated Learning for Arrhythmia Detection of Non-IID ECG. In: *Proc. International Conference on Computer and Communications (ICCC)*, IEEE, pp 1176–1180, DOI (10.1109/ICCC51575.2020.9344971), URL <https://doi.org/10.1109/ICCC51575.2020.9344971>

Zhou Y, Chen H, Li Y, Liu Q, Xu X, Wang S, Yap PT, Shen D (2021) Multi-task learning for segmentation and classification of tumors in 3D automated breast ultrasound images. *Medical Image Analysis* 70:101918, DOI (10.1016/j.media.2020.101918), URL <https://doi.org/10.1016/j.media.2020.101918>