Federated Learning Model Application to Gastric Cancer Diagnosis

Final Report

36105 iLab: Capstone Project

Group Members

David Bain (91082596)
Emmanuel Niko Sindayen (24888796)
James Murray (13879046)
Sinh Thanh Nguyen (25099704)
Taekjin Jeong (25099654)
Warisara Siriponpaiboon (25014616)

Abstract

Gastric cancer (GC), the fifth most common cancer globally, represents a significant public health concern. Although the data shows that the relative 5-year survival rate for this cancer is around 36%, the survival rate for early-stage GC exceeds 90%, underscoring the urgent need for faster, more accurate diagnostic methods. Convolutional Neural Networks (CNNs) have emerged as a promising tool in diagnostic imaging, helping to streamline the detection of GC. However, the usage of confidential patient data for third-party model training raises significant privacy concerns. As such, medical institutions often restrict access to data that is necessary to build robust, generalisable models.

To address this challenge, our team developed a federated learning model using FedAvg, simulating its effectiveness in detecting GC through histopathological images. The team trained three pre-built CNN models (ResNet, VGG-16, and VGG-19) and a custom model based on RESNET architechture as initial models and compared their performance before and after applying the federated learning framework. The observed results show improvements in accuracy and F1-score across four simulated clinics and one hold-out clinic after applying model weights through federated learning.

This project supports the potential of federated learning to enable model training while satisfying the privacy requirements which often prohibits direct access to sensitive data. Our findings demonstrated that even a simple implementation of federated learning can enhance the diagnostic accuracy of CNN models, facilitating faster, more reliable GC detection. The team hopes that the undertaking encourages medical institutions to explore federated learning as a viable solution for privacy-preserving, accurate cancer diagnostics.

Contribution Statement

Our group, taking project 7.2, with group members David Bain, Emmanuel Niko Sindayen, James Murray, Sinh Thanh Nguyen, Taekjin Jeong, and Warisarara Siriponpaiboon, state that each group member has contributed equally to complete the project during semester 2 2024.

Table of Contents

Abstract
Table of Contents 1
Introduction 2
Learning Algorithms 3
Transfer Learning 5
Federated Learning 6
Related Literature 8
Project Aims and Objectives 11
Project Questions 11
Methodology 12
Data collection 12
Data Preprocessing 12
Model Architecture 13
Training Configuration 15
Evaluation 16
Technical Implementation 18
Results 19
Clinical Model Results 19
Federated Learning Results 22
Discussion 24
Conclusion 25
References 26
Annex A – Model Development 31
ANEX B – Base Model Selection 36
Transfer Learning Process 36
Transfer Learning Results 37

Introduction

Gastric cancer (GC) is the fifth most common cancer worldwide by incidence and the fourth deadliest, with over 1 million new cases and 768,000 deaths globally in 2020. Projections estimate these numbers could rise to 1.77 million new cases and 1.27 million deaths by 2040, creating a growing burden on global health systems. Unfortunately, GC's early stages present symptoms indistinguishable from common gastrointestinal conditions such as ulcers or gastritis (Sheller et al., 2020). This lack of distinctive early symptoms often leads to postponed consultations and delayed diagnoses. Notably, the survival rate for GC is approximately 90% when detected early but drops sharply to around 30% when diagnosed at later stages.

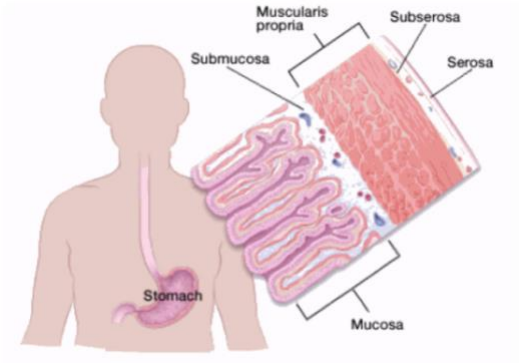


Figure 1. Gastric Cancer Detection

Multiple tests are conducted to confirm gastric cancer, including blood tests, ultrasound, and imaging tests like positron emission tomography (PET), computed tomography (CT), and magnetic resonance imaging (MRI). Among these, histopathology image analysis remains the gold standard for cancer diagnosis. Traditionally, pathologists conduct this analysis manually, meticulously screening tissue biopsies. However, this process is highly labour-intensive, time-consuming, and can be subjective. Despite the rigorous effort, diagnostic outcomes can vary depending on the pathologist's expertise and experience, introducing potential risks of misdetection or misdiagnosis due to human error. The ongoing shortage of skilled pathologists exacerbates this issue, leading to case backlogs and extended wait times for cancer detection (Yong et al., 2023).

Deep learning offers promising opportunities to standardise and automate histopathology analysis, potentially easing the diagnostic process. Several studies have explored deep learning models as diagnostic aids for pathologists (Deng et al., 2022). However, the development of these models is often hindered by restrictions on patient data, which are protected by medical institutions due to privacy concerns. This limitation prevents access to diverse datasets needed to create generalised algorithms that can reliably support pathologists across varied diagnostic scenarios.

One potential solution presented to address these privacy concerns is the use of Federated Learning for model training. The framework allows models to learn without direct access to the patients' confidential data. Instead, model-specific information (e.g, parameters or gradients) is aggregated across multiple institutions, enhancing the algorithms' generalised performance without compromising data privacy (Lu et al., 2020).

This project aims to validate the potential of federated learning to improve deep-learning classifier models for gastric histopathology. The team will develop an architecture that leverages the federated learning (FL) framework, allowing collaborative model training on distributed histopathology datasets without compromising data privacy. To evaluate the effectiveness of this framework, the team leveraged various learning algorithms and transfer learning. The methodology section provides a detailed discussion of the process.

Learning Algorithms

The team specifically employed three widely used deep learning models throughout the project as base models for gastric cancer detection: VGG-16, VGG-19, and ResNet-18.

- **VGG-16:** VGG stands for Visual Geometry Group. The term "deep" describes the number of layers, with VGG-16 or VGG-19 having 16 or 19 convolutional layers, respectively. Innovative object identification models are built using the VGG architecture. The VGGNet, created as a deep neural network, outperforms benchmarks on a variety of tasks and datasets outside of ImageNet. It also remains one of the most often used image recognition architectures today (Siddhesh, 2022). The model structure is characterised by 13 convolutional layers and 3 fully connected layers.

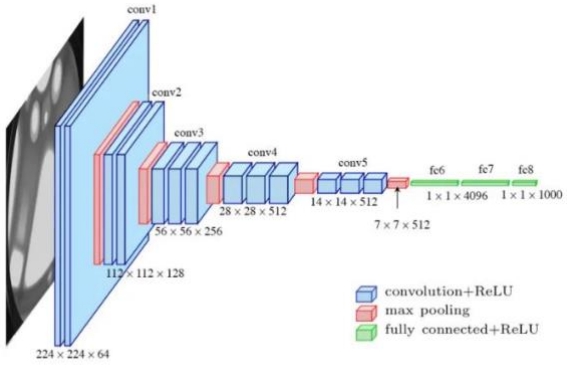


Figure 2. VGG-16

- **VGG-19:** VGG-19 was a follow-on from VGG-16 and has 16 convolutional layers and 3 fully connected layers. The extra layers offer some improvement in results but

come with an added computational overhead. VGG-16 and VGG-19 have been used in the production of medical imaging tasks due to their simplicity and performance.

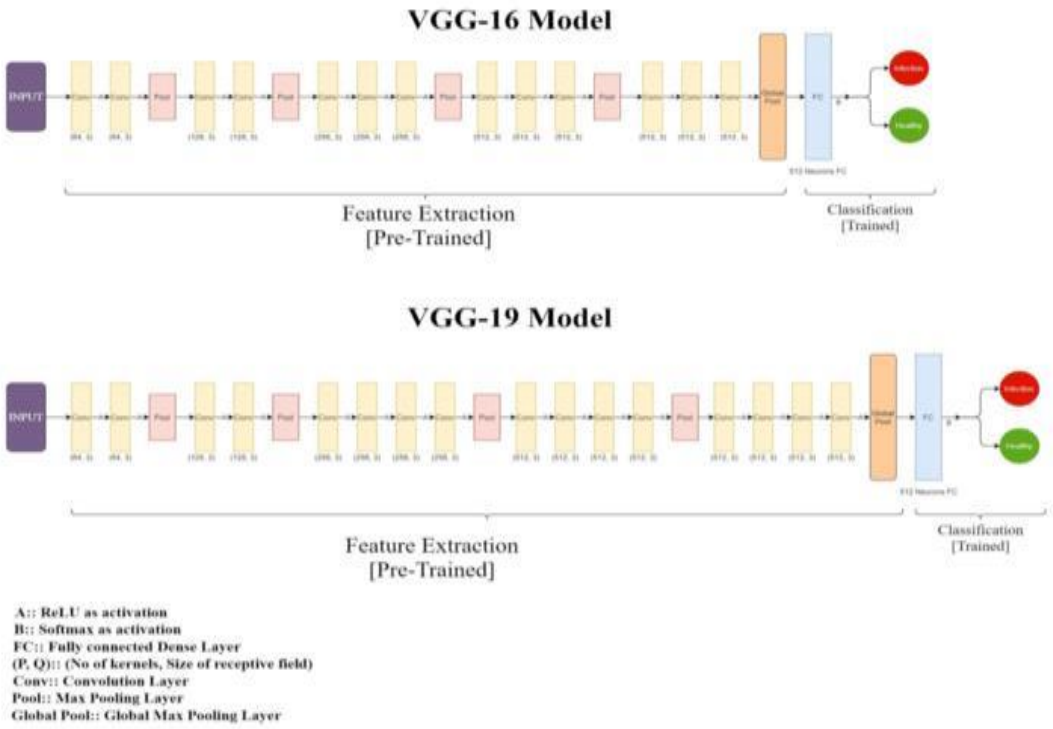


Figure 3. VGG-16 vs. VGG-19

- **RESNET-18:** Deep Residual Network or ResNet, developed by He et al. in 2016, is an architecture formed to defeat quandaries in deep learning training because deep learning training takes quite a lot of time and is limited to a certain number of layers. The defining feature of ResNet is the use of residual (or skip) connections, which bypass one or more layers. These connections allow the model to learn the identity function more efficiently and prevent the degradation of performance as the network depth increases. In conjunction with the skip connections, the' building' block approach allows the model to have deeper networks, and the deeper layer still learns features from the early layers. However, a key weakness of these models is their tendency to overfit. The model structure allows for all layers in the network to contribute to the classification task while keeping the model relatively small compared to the more extensive networks.

Transfer Learning

In this project, the team explored transfer learning by fine-tuning pre-trained state-of-the-art (SOTA) models identified earlier. This approach effectively addressed challenges related to limited data availability for training the algorithms, while conserving valuable time and computational resources (Kim et al., 2022).

Transfer learning is the application of knowledge learning from one domain to a different, adjacent domain, in a similar problem space (Deniz et al, 2018). In the Convolutional Neural Networks (CNN) context, this learning method involves two elements: feature extraction and fine-tuning (Saeed et al, 2023), with the fine-tuning of the model occurring in the last few layers. This implies that the feature extraction process was completed in an earlier step and that the features only need to be fine-tuned to the images in the problem space of the issue at hand (Deniz et al, 2023). Transfer learning can highlight the common layers (feature extractors) that remain the same and the altered classifier that is fine-tuned to adapt to the specific target task (Ahmad et al, 2023). This enables the model to leverage pre-learned features while tailoring its classification capabilities to new data or tasks, achieving improved accuracy and efficiency in scenarios with limited data.

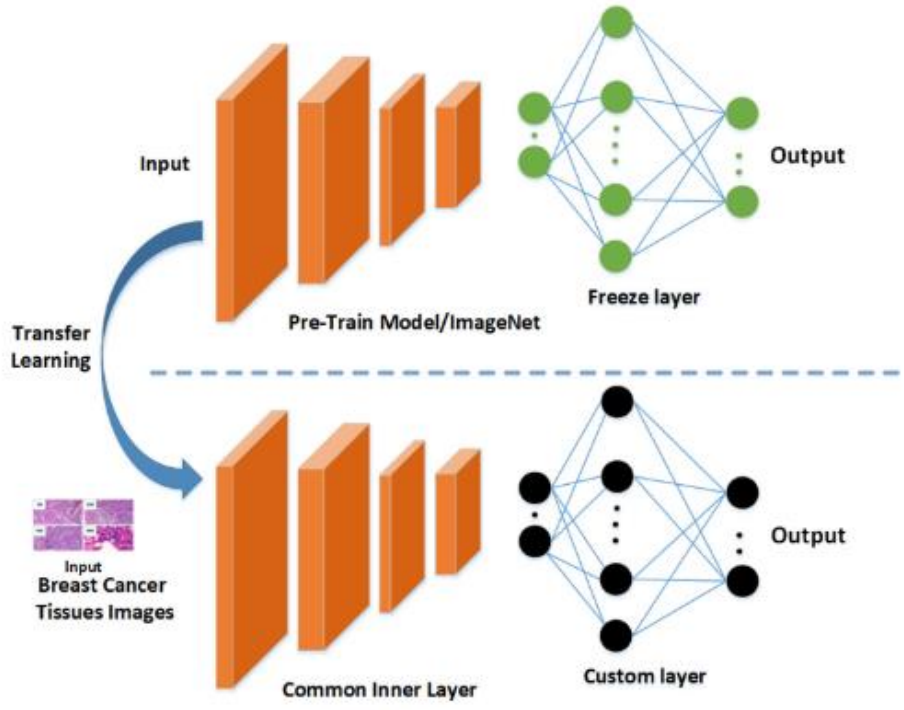


Figure 4. Transfer Learning

Transfer learning is widely utilised across industries employing computer vision, particularly in medical imaging. It has proven effective for image classification, matching the performance of fully trained models when hyperparameters and model usage are well understood (Victor et al., 2021). Leveraging pre-trained models to achieve SOTA results is a significant advantage especially considering the resources required to fully develop SOTA models for medical image classification tasks. The widespread adoption of transfer learning has led to numerous pre-trained models available in libraries like PyTorch and TensorFlow, most of which are trained on the ImageNet dataset.

ImageNet is a dataset comprising over 15 million labeled high-resolution images across approximately 22,000 categories, collected from the web and annotated using Amazon's Mechanical Turk. While ImageNet consists primarily of non-medical images, its low-level features (e.g., straight and curved lines) are universal to most image analysis tasks. Consequently, the transferred parameters (i.e., weights) from models trained using ImageNet provide a robust set of features, reducing the need for large datasets, as well as training time and memory costs (Sharma and Mehra, 2020), even in the medical domain.

Federated Learning

Federated Learning is a decentralised machine learning approach that enables multiple parties to collaboratively train a model without sharing their raw data. Instead of bringing all the data to a central server, federated learning allows each participant (such as hospitals, organisations, or devices) to train a local model on their own data and then send only the model updates (e.g., weights, gradients) to a central server.

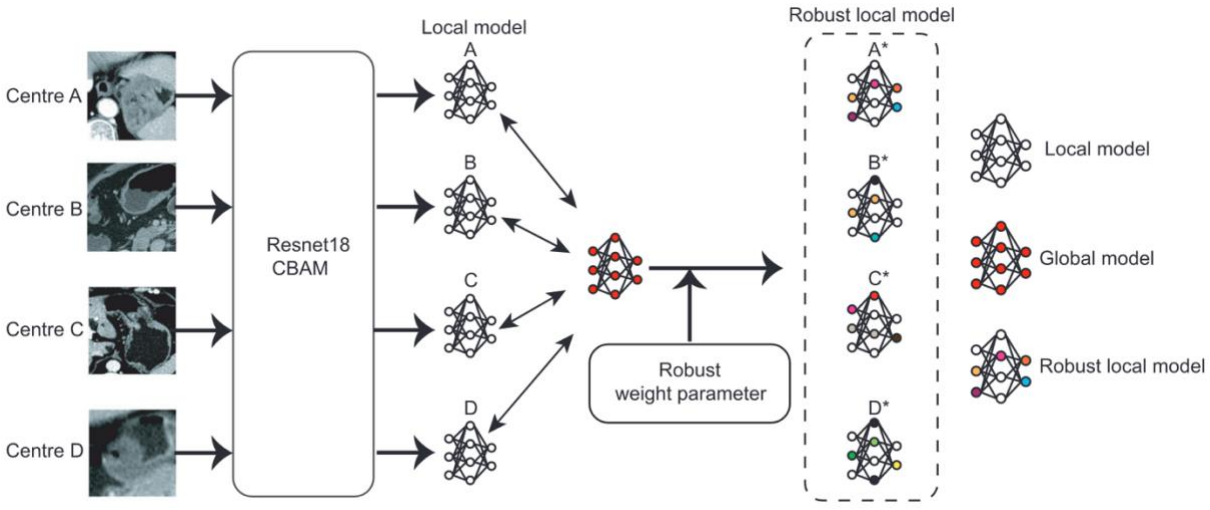


Figure 5. Supervised multiple instance learning in a federated framework

The term Federated Learning, first coined by researchers from Google in 2016 (Macmahon et al, 2016), was developed mostly in response to the 2012 White House paper that discussed a framework for protecting consumer data privacy while still promoting innovation in the global digital economy. The progressively pronounced ethical and moral requirements to preserve personal data, combined with advances in mobile technology spurred the increase in research into techniques such as FL that preserve privacy while still enabling the continuation of digital progress.

In the health domain, FL is gaining traction with the increasing health data privacy protection requirements (Teo et al, 2024), with the most potential benefit seen from the use of FL in medical imaging and neural networks due to the data requirements associated with the development of neural networks in support of image classification and segmentation tasks. This approach allows institutions to retain control over sensitive patient information while still benefiting from collective insights drawn from larger, diverse datasets across multiple institutions. Consequently, federated learning supports the effective deployment of predictive models in clinical environments, leading to potentially faster diagnoses, tailored treatments, and improved patient outcomes. Additionally, by democratising access to data for researchers, federated learning promotes innovation, accelerates model development, and reduces costs associated with traditional centralised data processing. Through its capacity to leverage large-scale collaborative datasets, federated learning is reshaping the healthcare technology landscape, making it possible to advance patient care through secure, data-driven insights on a global scale (Prayitna et al, 2021).

Through FL, machine learning models used by medical institutions would have the following positive characteristics:

- No requirement to share data
- Control over engagement with the system
- Improved models for deployment in clinical settings
- Potentially faster and better patient outcomes
- Increased accessibility for researchers of all natures
- Reduced costs associated with development of ML
- Access to more data to build better models

Thus, FL has the potential to play a pivotal role in healthcare by enabling the development of advanced machine learning models while preserving patient privacy and data security, critical in clinical settings where confidentiality is paramount.

Related Literature

The potential of convolutional neural network (CNN) models in diagnosing diseases, such as gastric cancer, has been a consistent source of deep learning research. Notable examples are Hirasawa et al's (2018) robust Single Shot MultiBox Detector CNN diagnostic system which successfully identified 98.6% of lesions with a diameter of at least 6 mm with the incorrect tagging noted as unique edge cases, Ikenoyama et al's (2021) Inception-v3 CNN, which detected early gastric cancer (EGC) cases more rapidly than endoscopists, and Tang et al's (2020) Darknet-53 CNN, which were found to enhance clinicians' diagnostic capabilities.

In the project's scope of histopathological image analysis, Yoshida et al. (2018) explored e-Pathologist, the first automated image analysis software for aiding gastric cancer diagnosis. This software showed promising potential, with anticipated improvements in its screening capabilities for broader clinical application. Expanding on this work, Song et al. (2020) developed an AI system to assist pathologists in analyzing gastric histopathology images. This system could flag slides that required re-examination or further testing, functioning as both a pre-analytical prioritisation tool and a second-opinion system, especially valuable in complex cases. A systematic review by Klang et al. (2023) on the application of deep learning in gastric cancer (GC) detection highlighted the effectiveness of AI systems in enhancing diagnostic accuracy and image segmentation. However, the review also identified significant challenges, such as the need for robust model performance across diverse populations and varied imaging conditions, which would necessitate larger, more heterogeneous datasets. Issues related to the validation and standardisation of AI models were flagged as critical barriers to clinical adoption. Klang et al. noted that these challenges may have been exacerbated by the limitations of the studies reviewed, as most were conducted in single-center settings, which could restrict generalizability.

To address data availability and privacy concerns, Deng et al. (2022) proposed federated learning (FL) as an effective solution. FL enables multiple institutions to collaborate on model training without exchanging sensitive patient data, thereby enhancing data privacy and security while supporting model validation and standardisation across sites. In a comparative study, Sheller et al. (2020) evaluated FL alongside other collaborative learning methods, such as Institutional Incremental Learning (IIL) and Cyclic Institutional Incremental Learning (CIIL). They found that FL achieved superior outcomes, delivering the highest rate of model improvement per epoch and yielding better-performing models on average.

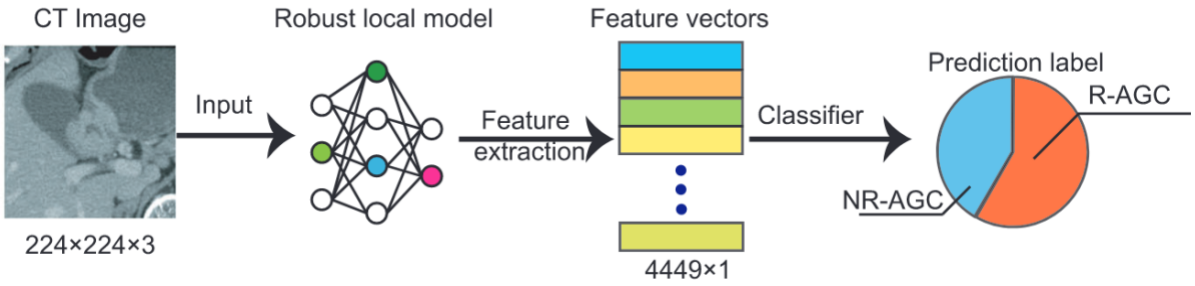


Figure 6. CT image to feature extraction and classification

Feng et al. (2024) developed a robust federated learning model (RFLM) designed to identify high-risk patients for postoperative gastric cancer recurrence. Their model reportedly enhanced local hospital diagnostic models and boosted performance when applied to a separate public lung cancer dataset, highlighting FL's adaptability. However, the study's peer review raised methodological concerns, particularly in the identification of regions of interest (ROI), sample sizes, and the privacy protection mechanisms for generated data.

Other FL applications have also demonstrated promise in improving diagnostic imaging and patient outcomes. For instance, Pati et al. (2022) employed FL in diagnosing common, fatal brain tumors, with positive effects on the final consensus model's performance. Almufareh et al. (2023) noted significant advancements in breast cancer diagnosis through FL by consolidating data from multiple institutions without compromising patient privacy. Similarly, Lu et al. (2020) used FL to develop models for whole-slide histopathological images, enabling broader institutional data contributions and fostering models that generalise effectively to unseen data.

Despite these advantages, FL has limitations, particularly its vulnerability to security threats such as backdoors, data poisoning, membership inference, GAN-based attacks, and differential privacy breaches (Hasan, 2023). Additionally, even model weight transfers may risk exposing sensitive information through reverse engineering (Sheller et al., 2023). Bias propagation also poses a challenge, potentially leading to unequal treatment outcomes across patient groups (Chang & Shikori, 2023). However, FL remains a transformative approach in healthcare AI, supporting the development of more accurate diagnostic models without compromising data confidentiality. Mitigation strategies, like differential privacy, can further reduce these risks.

In terms of the dataset, GasHisSDB (Hu et al.) was developed with images sourced from Longhua Hospital Shanghai University of Traditional Chinese Medicine and augmented by biomedical researchers from Northeastern University and pathologists from Liaoning Cancer Hospital and Institute. The dataset's usability was validated through classification using classical machine learning methods (Random Forest, linear SVM) and deep learning

models (VGG16, ResNet50, ViT), all of which performed competently. Additional studies by Yong et al. (2023) and Khayatian et al. (2024) further supported the dataset's effectiveness for training gastric image classifiers, establishing its viability in deep learning research for gastric cancer.

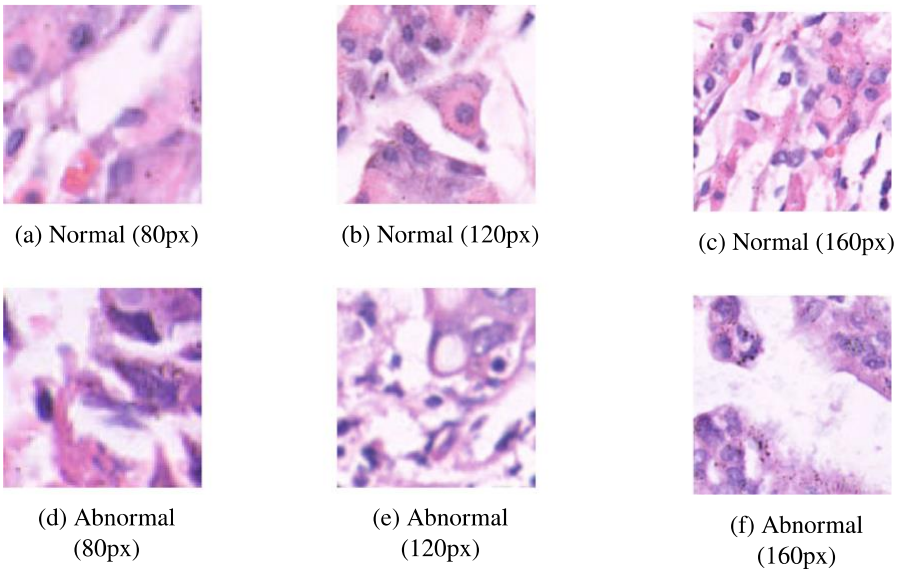


Figure 7. Sample of normal and abnormal tumour types in all sizes

The design of this project closely aligns with the study by Feng et al., in which the authors developed a federated learning framework, the Robust Federated Learning Model (RFLM), and applied it to local models trained on data from four separate centers. Feng et al. assessed model performance both before and after implementing the RFLM and compared it with other federated learning frameworks like FedAvg and FedProx. Unlike Feng et al., who focused on CT images from proprietary sources, this project will use a publicly available histopathological image dataset (GasHisSDB). This decision not only enhances the transparency of model predictions but also emphasises the unique insights derived from histopathological image analysis, a recognised gold standard in gastric cancer diagnosis.

Furthermore, similar to Hu et al. (2022), this project will employ VGG and ResNet models to classify the GasHisSDB dataset. However, this project will assess the performance of the classifier models by simulating five distinct clinical settings (four contributor clinics and one held-out) by partitioning the dataset, then applying the federated learning framework to enable locally informed model updates. This approach highlights federated learning's potential to address data privacy concerns while underscoring the diagnostic effectiveness of VGG and ResNet in the context of medical imaging analysis. A detailed discussion of the team's approach can be found in the methodology section.

Project Aims and Objectives

The objective of this project is to develop a predictive tool for faster intervention and treatment of potential gastric cancer diagnoses. The project focused on the histological analysis of gastrointestinal tissues, leveraging advanced image classification techniques and Federated Learning to ensure data privacy. The emphasis on Federated Learning enables collaboration between multiple institutions without sharing sensitive patient data, offering potential applications beyond gastric cancer diagnosis to other clinical contexts. The tool aims to aid in early detection, improving treatment outcomes while maintaining compliance with data privacy regulations.

Project Questions

- Can a single image classification model trained on a small dataset (single clinic) outperform the same model developed across multiple datasets (multiple clinics) using Federated Learning to ensure data privacy?
- Can Federated Learning enhance the predictive performance of models trained in different clinical settings?

Methodology

Data collection

The GasHisSDB dataset contains 245,196 sub-size images in two categories, including 97,076 abnormal images and 148,120 normal images. There are three sub-databases, 160x160 pixels, 120x120 pixels, and 80x80 pixels. Figure 8 identifies a correlation in the reduced number of available images as the resolution increases. There are only 33,284, or nearly 78% less images with 160x160 pixel resolution compared to 146,651 images with 80x80 pixels.

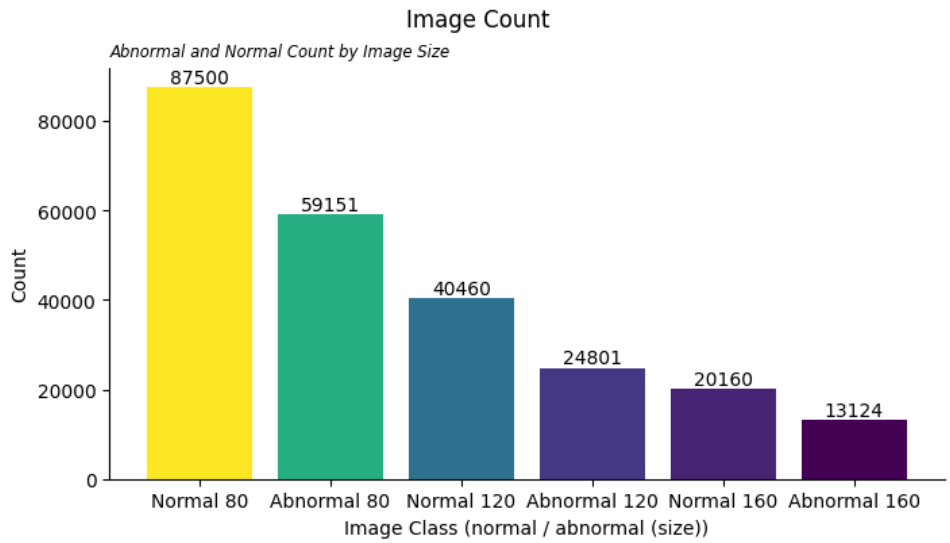


Figure 8. GasHisSDB dataset profile

To examine the robustness of Federated Learning method the dataset was divided into five chunks, simulating five clinical institutions.

Data Preprocessing

Image Selection

Figure 9 shows the variance in the images identifying a number of low contrast images which make it challenging in machine learning, even with pre-trained SOTA models used in transfer learning.

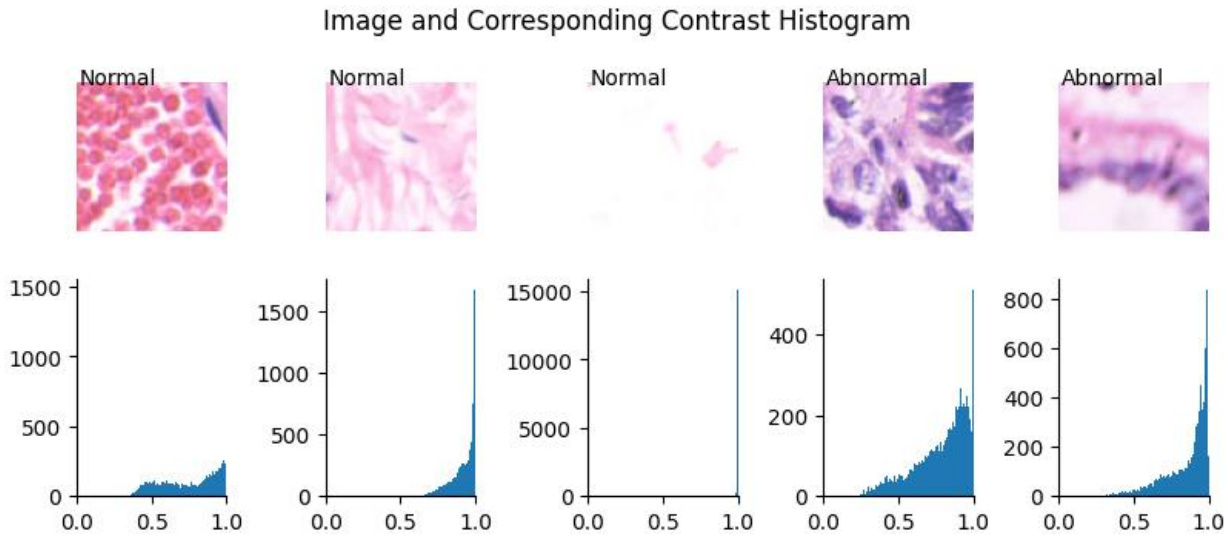


Figure 9. histopathological image variances

To emulate *real world* conditions, the choice was made to *not* remove any of the extremely low contrast images from the data set. Even though studies have found that removing or employing contrast adjustment augmentation techniques improves the accuracy of the model (Shorten & Khoshgoftaar, 2019).

Resizing and Normalisation

In the data preprocessing phase of our project, a crucial step involved resizing and normalising the histopathological images to ensure compatibility with the pre-trained models. Models such as ResNet18, VGG16, and VGG19 are pre-trained using ImageNet which determines their architecture to **224x224** inputs. Furthermore, this step helps the models to converge faster during training and improves overall performance. Therefore, the 160x160 pixel images imported from each clinic were resized to **224x224**. We used the normalisation parameters derived from the ImageNet dataset, which are mean values of **[0.485, 0.456, 0.406]** and standard deviation values of **[0.229, 0.224, 0.225]** for the RGB channels. After normalisation, the dataset corresponding to each clinic was split into two datasets; train and validation with a ratio of 80:20.

Model Architecture

Base Models Selection

Before experimenting with Federated Learning model, the team has discovered several base CNN models with different settings (Anex B) and decided to proceed further with three pre-trained models, VGG16, VGG19, and ResNet18, and a full-trained ResNet18 with additional dropout layers. Regarding the three pre-trained models, the classifiers were modified for use in a binary classification problem. Moreover, several drop-out layers were

also added to avoid overfitting. All parameters in a pre-trained model except for ones in classifier layer are “frozen” to utilise the effectivity of pre-trained weights. In other words, only classifier layers were trained, and the convolutional layers’ parameters were unchanged.

Federated Learning Framework

Due to time, computational and ability constraints, this project decided to employ Federated Averaging (FedAvg) method as the Federated Learning model. It averages model weights from local clients after each round of training, which is computationally less expensive compared to other complex federated learning algorithms. This simplicity ensures faster implementation and easier debugging, making it an ideal choice for initial federated learning experiments.

In this set up, the global model was initialised with a set of parameters, corresponding to one of the base models. Four of five clients will then update their local models with these parameters and train them using the local data. After training, local clients will extract the updated parameters and send them to the central server, where FedAvg aggregates the updates by averaging the weights. The refined global model was redistributed to all participating clients for the next training round. This process can be visualised in Figure 10.

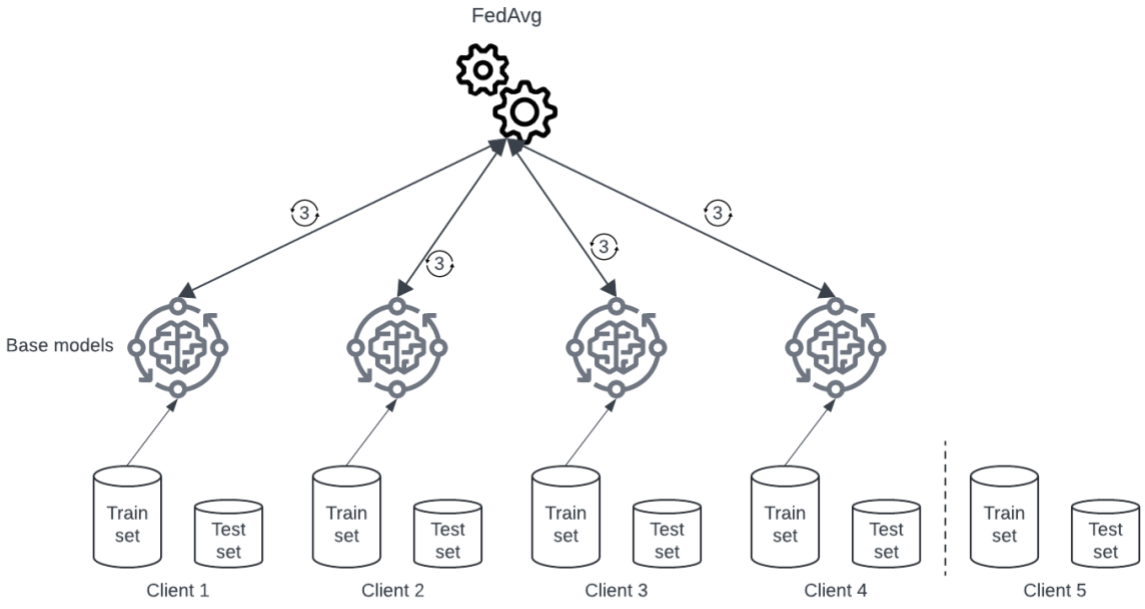


Figure 10. Federated Learning

Training Configuration

Training Setup

Four clinical institutions were involved in three round training process. Each training round requires each clinic to train its local model using its train set for 20 epochs with a batch size of 32.

Loss Function

This project utilised Binary Cross-Entropy Loss (BCELoss) as the primary loss function. This choice was driven by the binary nature of the gastric cancer classification problem, where each histopathological image is classified as either positive or negative for gastric cancer. The function's formula is described as follow:

$$BCE = -\frac{1}{N}\sum_{i=0}^{N} y_i \cdot log(\hat{y}_i) + (1 - y_i) \cdot log(1 - \hat{y}_i)$$

Figure 11. FedAvg Training

Optimisation Algorithm

The Adam optimiser was selected for its ability to efficiently handle large-scale datasets and complex models, like those used in image-based cancer detection. Unlike standard gradient descent methods, Adam adapts to learning rate for each parameter, which allows for faster convergence and model stable training. This adaptive nature minimises the need for extensive hyperparameter tuning, making it particularly advantageous in computational resources and time-limited scenarios. In experimentations, the learning rate of 0.001 was initialised for the optimiser.

Regularisation Techniques

Apart from adding drop-out layers, we considered employing Learning Rate Decay and Early Stopping as regularisation techniques. Learning rate decay was employed to reduce the learning rate as training progressed gradually. This approach helps in stabilising the convergence process. Initially, a higher learning rate facilitates rapid learning, but as the model approaches the optimal solution, a lower learning rate prevents overshooting and ensures fine-tuning. This adjustment improves the models' accuracy and helps avoid oscillations near the minimum loss. In this project, the learning rate is divided by 10 after every 2 epochs without reducing in the loss. Early stopping was integrated into the training process to prevent overfitting. By monitoring the model's performance on a validation set,

training was halted once the validation loss stopped improving, even if the maximum number of epochs had not been reached. This method conserves computational resources and ensures that the model generalises well to unseen data. Early stopping was triggered after 4 epochs without model improvement in the project.

Evaluation

Evaluation Methodology

Method 1: Accuracy - Federated Model Performance Across Contribution Clinics

In the first evaluation method, we aimed to assess the efficacy of the FedAvg model across the participating clinics that contributed their data for training. Initially, the model was trained locally on the datasets of four simulated clinics, followed by aggregation of the local updates through FedAvg approach to generate a global model. However, before making predictions, the global model was fine-tuned on each clinic's local training data to better adapt the generalised knowledge to the specific nuances of each clinic's dataset. This fine-tuning step was crucial to align the global model with the unique characteristics of each clinic's data. Once the model was fine-tuned, it was applied to the corresponding test datasets of the same clinics to evaluate its performance.

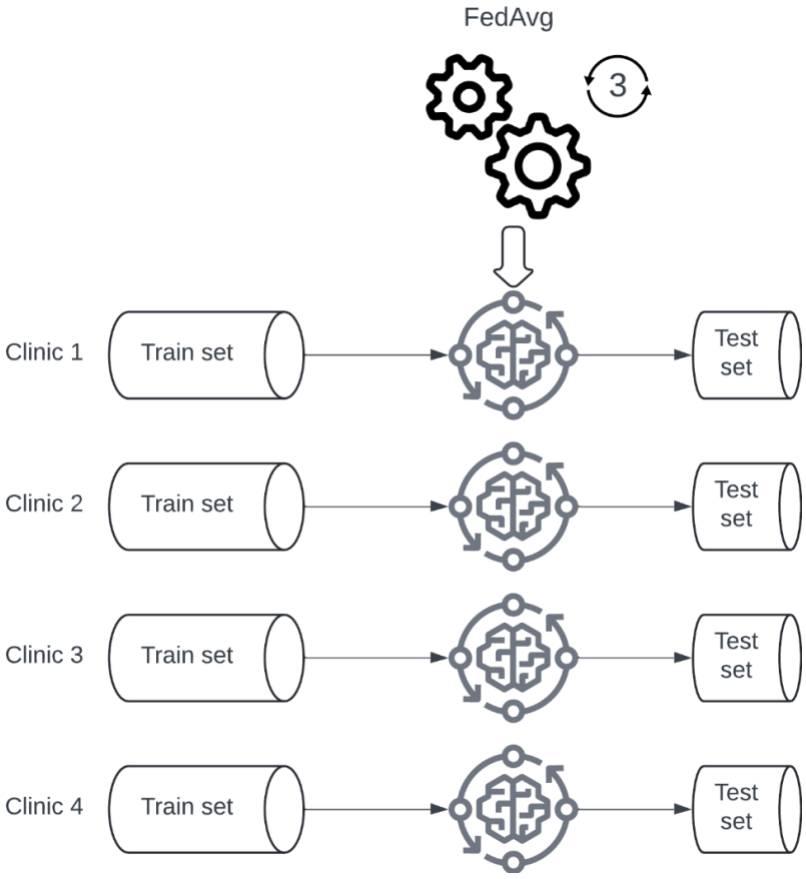


Figure 12. FedAvg Training

This evaluation method is critical as it demonstrates the immediate benefits of federated learning for the participating clinics. By comparing the model's predictions to the ground truth labels in the test datasets, we could gauge the performance improvements resulting from collaborative learning. This method highlights whether the federated approach effectively generalises the knowledge across multiple clinics, thereby enhancing the predictive accuracy of local models. The insights gained here are crucial for validating the model's utility in real-world scenarios where different institutions pool their learning without sharing sensitive data.

Method 2: Scalability - Assessing Federated Model Utility for Late-Joining Clinics

The second evaluation scenario was designed to simulate a clinic that joins the federated learning network after the initial training rounds have already concluded. The federated model trained using data from the first four clinics, was fine-tuned using the fifth clinic's local training data before making predictions on its test set. This fine-tuning step allowed the global model to adapt to the specific characteristics of the late-joining clinic's dataset.

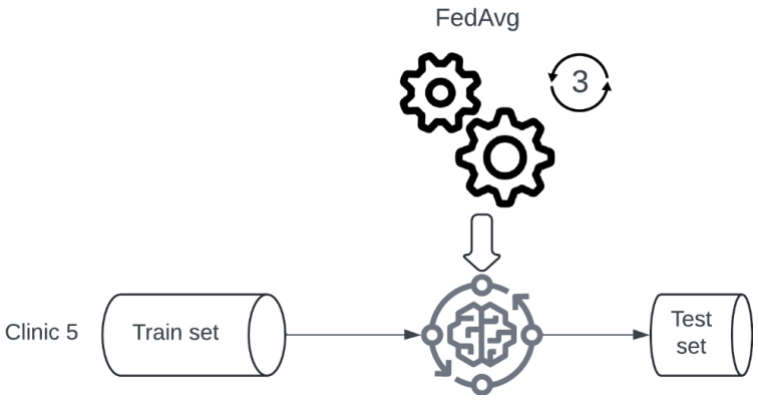


Figure 13. FedAvg Clinic 5 introduction

This method addresses a significant question in federated learning: Can a clinic benefit from the collective knowledge of the federated model even if it joins the process late? By testing the fine-tuned global model on the fifth clinic's test data, we assessed the model's ability to generalise its learning and deliver accurate predictions for institutions that join the federated network at a later stage. This scenario underscores the flexibility and scalability of the federated learning approach, ensuring that even new participants can quickly achieve high diagnostic performance without the need for extensive initial training.

Evaluation Metrics

We selected Area Under the Receiver Operating Characteristic Curve (AUROC) and F1 score as the primary metrics to evaluate the performance of our models. These metrics were chosen due to their effectiveness in capturing nuances of medical diagnosis,

particularly in a critical field such as gastric cancer detection. In the context of gastric cancer detection, where false negatives can have severe consequences, AUROC is particularly valuable. It allows us to understand how well the model separated cancerous from non-cancerous cases without committing to a specific threshold. A higher AUROC score indicates that the model can distinguish between the two classes more effectively, crucial for ensuing reliable predictions in medical applications.

The F1 score, defined as the harmonic mean of Precision and Recall, is particularly suited for scenarios where balance between false positives and false negatives is critical. In the diagnosis of gastric cancer, precision ensures that when the model predicts cancer, it is likely correct, minimising unnecessary stress and treatments. Conversely, recall ensures that most of the actual cancer cases are identified, minimising the risk of missed diagnoses. By focusing on this metric, we ensure that the model not only achieves high accuracy but also maintains a meaningful balance between correctly identifying positive cases and minimising false alarms.

$$\boxed{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

$$\boxed{Recall} = \frac{TP}{TP + FN} \quad \text{Known as True Positive Rate}$$

$$\boxed{Specificity} = \frac{TN}{FP + TN} \quad \text{Known as True Negative Rate}$$

$$\boxed{Precision} = \frac{TP}{TP + FP}$$

$$\boxed{F1 - Score} = 2\frac{Precision * Recall}{(Precision + Recall)}$$

Figure 14. Metrics formula

Technical Implementation

This project was implemented using PyTorch and Torchvision for base model or clinical development and aggregated using federated learning. The base model selection phase was conducted on a Nvidia GeForce RTX 4060 8GB integrated local hardware and the employment of the Federated Learning method on Kaggle's GPU's to allow extensive training without cloud costs. Challenges such as client-server synchronisation and efficient weight aggregation were addressed through customised implementations with the federated learning framework.

Results

Clinical Model Results

Results of the individual clinics using pre-trained models highlighted in Table 1 constituted our baseline. They were relatively consistent, each within a range of 93-96% AUROC, and similarly the F1 scores were between 81-84%. The result highlights a trend of pre-trained models to overfit and therefore the number of epochs needed to be capped at 10. However, the CustomResNet18 model for individual clinics performed distinctly worse than the pre-trained models, achieving between 82-92% AUROC and 40-75% F1.

Table 1. Clinical model results

Evaluation Method 1			
Model	**Clinic**	**AUROC**	**F1**
VGG16	Clinic_0	0.9421	0.8386
	Clinic_1	0.9352	0.8315
	Clinic_2	0.9295	0.8156
	Clinic_3	0.9421	0.8322
VGG19	Clinic_0	0.9506	0.8414
	Clinic_1	0.9416	0.8273
	Clinic_2	0.9414	0.8189
	Clinic_3	0.9474	0.8397
ResNet18	Clinic_0	0.95	0.8319
	Clinic_1	0.9485	0.8317
	Clinic_2	0.9445	0.8142
	Clinic_3	0.9601	0.8437
CustomResNet18	Clinic_0	0.9261	0.7527
	Clinic_1	0.8958	0.6145
	Clinic_2	0.8803	0.7274
	Clinic_3	0.8261	0.4069

It highlights the trade-off between large pre-trained models, which all have over 100 million weights and were trained on ImageNet with over 14 million images. Compared to our custom Resnet model which has only 11 million weights and was trained on only 30,000 images. Whilst the custom Resnet model performed worse at the individual clinic level, its strength is revealed in a federated learning context. Reviewing False Positive and False Negatives in Figure 15. VGG19 pre-trained confusion matrix, the clinical model outlines 310

patients requiring clinician review. It emphasises the need to consider the probability confidence level before the sigmoid function and expand this clinical review to all results within a 30-70% probability range.

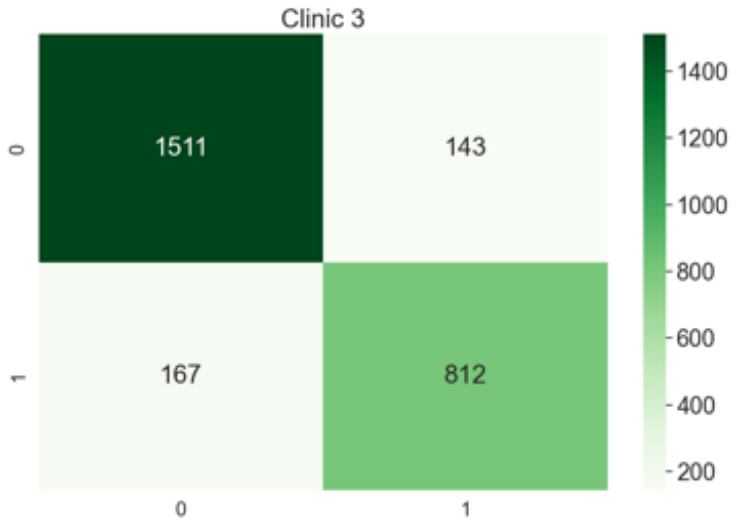


Figure 15. VGG19 pre-trained confusion matrix

When introducing a new clinic, 'Clinic 5' in Table 2, we see VGG16, VGG19 and Resnet18 in all performing similarly to the training clinics, with 92-94% AUROC and 81-83% F1 score, validating successful introduction of new hospitals without retraining. Custom Resnet18 performed at the upper end of the custom model range for Clinic 5 with 90% AUROC and 77% F1, and predictably below the pre-trained models.

Table 2. Clinic 4 result

Evaluation Method 2			
Model	**Clinic**	**AUROC**	**F1**
VGG16	Clinic_4	0.9283	0.8164
VGG19	Clinic_4	0.9421	0.8396
ResNet18	Clinic_4	0.9357	0.8257
CustomResNet18	Clinic_4	0.9081	0.7697

Comparatively in Figure 16, we notice improved True Positive results, yet slightly worse True Negative(TN) and False Negative(FN) results. This reinforces the need for a larger training set.

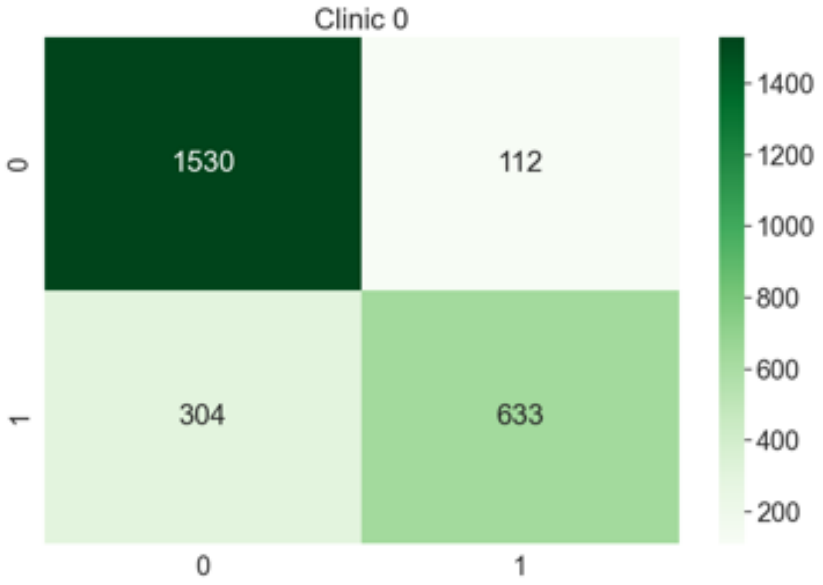


Figure 16. Custom Resnet model

Federated Learning Results

Table 3 shows an improvement in results against each pre-trained clinic of 94-96% AUROC or 1% improvement, and 82-86% F1 or 2% improvement. It indicates that federated learning has marginally improved individual clinical performance, yet more importantly, it has improved the precision and recall thus reducing false negatives and false positives (FP). The FedAvg CustomResNet18 model has improved significantly to 97% AUROC and 89-90% F1, notably 5% and 15%, respectively, over its baseline, and additionally outperforms all pre-trained models in terms of accuracy and sensitivity.

Table 3. Federated Learning with different clinical models

Evaluation Method 1			
Model	**Clinic**	**AUROC**	**F1**
FedAVG VGG19	Clinic_0	0.9564	0.8518
	Clinic_1	0.9513	0.8432
	Clinic_2	0.9538	0.8545
	Clinic_3	0.9595	0.8608
FedAVG VGG16	Clinic_0	0.9524	0.8588
	Clinic_1	0.9504	0.8424
	Clinic_2	0.9462	0.8242
	Clinic_3	0.9588	0.8523
FedAVG ResNet18	Clinic_0	0.9529	0.8446
	Clinic_1	0.9541	0.8462
	Clinic_2	0.9499	0.8260
	Clinic_3	0.9653	0.8589
FedAVG CustomResNet18	Clinic_0	0.9734	0.8973
	Clinic_1	0.9649	0.8855
	Clinic_2	0.9737	0.9006
	Clinic_3	0.9733	0.9029

Table 4 shows consistent performance across the pre-trained models 94-95% AUROC, an improvement of around 2%, and 82-85% F1 a similar improvement of 2% against the individual clinical results.

Notably, FedAvg CustomResNet18 model displays a significant improvement for clinic 5 of 5% AUROC and 9% F1 score over its baseline clinical counterpart and performs slightly better than the FedAvg pre-trained models.

Table 4. Federated Learning results with Clinic 5

Evaluation Method 2			
Model	**Clinic**	**AUROC**	**F1**
FedAvg VGG19	Clinic_4	0.9514	0.8529
FedAvg VGG16	Clinic_4	0.9419	0.8247
FedAvg ResNet18	Clinic_4	0.9461	0.8407
FedAvg CustomResNet	Clinic_4	0.9506	0.8585

Discussion

Background research highlighted a focus on using transfer learning with pre-trained models to create the baseline for a federate learning (FL) framework, with the expectation that large models trained on large datasets would prove more resilient to poor images and deliver a more robust overall performance. The results of this project prove this behaviour was consistent at the individual clinic level with a slightly improved FL performance of around 2%.

The characteristics of the custom Resnet model with the dropout in the convolutional layers in the federated setting, offered a superior model than during training in a single clinic. This is reflected by as much as 15% improvement for the F1 score. Additionally, we note that the custom Resnet model also marginally outperformed the larger pre-trained models for both accuracy and sensitivity by around 2%.

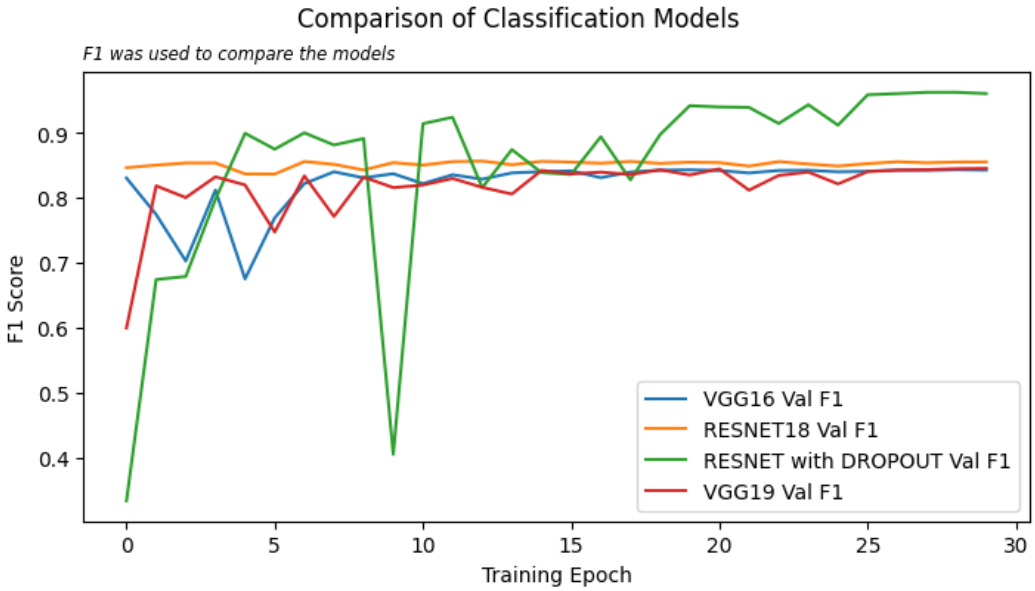


Figure 17. Model Training Comparison

Figure 17 shows that it took longer for the custom Resnet model to reach better results than the other pre-trained models when trained on a single clinic, however it proved to be more resilient to insufficient data and other anomalies in the clinical setting. We anticipate that with more training epochs and more images/hospitals in the federated setting this model will continue to improve beyond the current results and surpass the other pre-trained model for this complex classification task. Based on our research, this kind of custom model has yet to be deployed in a federated setting. It presents a significant opportunity for future research in model design for federated settings.

Conclusion

Early detection of gastric cancer is essential due to the disease's high mortality rate and diagnostic challenges. Timely identification can drastically improve patient outcomes, with early-stage detection offering a five-year survival rate of up to 90%. Our research aims to support faster, more accurate diagnoses, potentially saving lives by improving biopsy screening and accelerating detection times.

A key challenge in this area is maintaining patient data privacy, as hospitals handle sensitive information that, if breached, could lead to severe consequences, including privacy violations, identity theft, and regulatory repercussions. Federated learning addresses these concerns by enabling collaborative model training across institutions without compromising individual data confidentiality. This privacy-preserving approach has become especially relevant for healthcare applications.

Our study demonstrates the effectiveness of the FedAvg approach when applied to various deep learning architectures including VGG19, VGG16, ResNet18, and CustomResNet18 in detecting gastric cancer. FedAvg consistently enhanced F1 and AUROC scores, showing significant gains in predictive accuracy across different clinical settings. Notably, our custom ResNet model showed remarkable improvement in predictive performance after applying FedAvg, mainly due to its federated learning-enabled stabilisation of convergence and parameter sharing.

Resource constraints limited our ability to compare the custom ResNet with fully trained VGG models. However, our findings reinforce that a single isolated model cannot outperform a federated approach, affirming that federated learning enhances predictive performance across diverse settings.

Future research could further enhance model performance by fully training VGG models, expanding datasets with additional gastric images, and exploring alternative federated learning methodologies to continue advancing early gastric cancer diagnosis.

References

"Efficient Gastrointestinal Disease Classification Using Pretrained Deep Convolutional Neural Network - ProQuest." n.d. Accessed November 2, 2024. https://www.proquest.com/docview/2799630724?accountid=17095&parentSessionId=1mlLnft2oPFQ4Hr9bP4KD3%2FtsGTS6IMRrRMBw03jC%2Fk%3D&pq-origsite=primo&sourcetype=Scholarly%20Journals.

"Models and Pre-Trained Weights — Torchvision 0.20 Documentation." Accessed October 23, 2024. https://pytorch.org/vision/stable/models.html.

Ahmad, Nouman, Sohail Asghar, and Saira Andleeb Gillani. 2022. "Transfer Learning-Assisted Multi-Resolution Breast Cancer Histopathological Images Classification." *The Visual Computer* 38 (8): 2751–70. https://doi.org/10.1007/s00371-021-02153-y.

Almufareh, Maram Fahaad, Noshina Tariq, Mamoona Humayun, and Bushra Almas. 2023. "A Federated Learning Approach to Breast Cancer Prediction in a Collaborative Learning Framework." *Healthcare* 11 (24): 3185. https://doi.org/10.3390/healthcare11243185.

Chang, Hongyan, and Reza Shokri. 2023. "Bias Propagation in Federated Learning." arXiv. https://doi.org/10.48550/arXiv.2309.02160.

Deng, Yang, Hang-Yu Qin, Yan-Yan Zhou, Hong-Hong Liu, Yong Jiang, Jian-Ping Liu, and Ji Bao. 2022. "Artificial Intelligence Applications in Pathological Diagnosis of Gastric Cancer." *Heliyon* 8 (12): e12431. https://doi.org/10.1016/j.heliyon.2022.e12431.

Deniz, Erkan, Abdulkadir Şengür, Zehra Kadiroğlu, Yanhui Guo, Varun Bajaj, and Ümit Budak. 2018. "Transfer Learning Based Histopathologic Image Classification for Breast Cancer Detection." *Health Information Science and Systems* 6 (1): 18. https://doi.org/10.1007/s13755-018-0057-x.

Feng, Bao, Jiangfeng Shi, Liebin Huang, Zhiqi Yang, Shi-Ting Feng, Jianpeng Li, Qinxian Chen, et al. 2024. "Robustly Federated Learning Model for Identifying High-Risk Patients with Postoperative Gastric Cancer Recurrence." *Nature Communications* 15 (January):742. https://doi.org/10.1038/s41467-024-44946-4.

Hasan, Jahid. 2023. "Security and Privacy Issues of Federated Learning." arXiv. https://doi.org/10.48550/arXiv.2307.12181.

Hirasawa, Toshiaki, Kazuharu Aoyama, Tetsuya Tanimoto, Soichiro Ishihara, Satoki Shichijo, Tsuyoshi Ozawa, Tatsuya Ohnishi, et al. 2018. “Application of Artificial Intelligence Using a Convolutional Neural Network for Detecting Gastric Cancer in Endoscopic Images.” *Gastric Cancer* 21 (4): 653–60. https://doi.org/10.1007/s10120-018-0793-2.

Hu, Weiming, Chen Li, and Xiaoyan Li. n.d. “GasHisSDB: A New Gastric Histopathology Image Dataset for Computer Aided Diagnosis of Gastric Cancer.”
Ikenoyama, Yohei, Toshiaki Hirasawa, Mitsuaki Ishioka, Ken Namikawa, Shoichi Yoshimizu, Yusuke Horiuchi, Akiyoshi Ishiyama, et al. 2021. “Detecting Early Gastric Cancer: Comparison between the Diagnostic Ability of Convolutional Neural Networks and Endoscopists.” *Digestive Endoscopy* 33 (1): 141–50. https://doi.org/10.1111/den.13688.

Khayatian, Danial, Alireza Maleki, Hamid Nasiri, and Morteza Dorrigiv. 2024. “Histopathology Image Analysis for Gastric Cancer Detection: A Hybrid Deep Learning and Catboost Approach.” *Multimedia Tools and Applications*, August. https://doi.org/10.1007/s11042-024-19816-2.

Kim, Hee E., Alejandro Cosa-Linan, Nandhini Santhanam, Mahboubeh Jannesari, Mate E. Maros, and Thomas Ganslandt. 2022. “Transfer Learning for Medical Image Classification: A Literature Review.” *BMC Medical Imaging* 22 (1): 69. https://doi.org/10.1186/s12880-022-00793-7.

Klanecek, Zan, Tobias Wagner, Yao-Kuan Wang, Lesley Cockmartin, Nicholas Marshall, Brayden Schott, Ali Deatsch, et al. “Uncertainty Estimation for Deep Learning-Based Pectoral Muscle Segmentation via Monte Carlo Dropout.” *Physics in Medicine & Biology* 68, no. 11 (May 2023): 115007. https://doi.org/10.1088/1361-6560/acd221.

Klang, Eyal, Ali Sourosh, Girish N. Nadkarni, Kassem Sharif, and Adi Lahat. 2023. “Deep Learning and Gastric Cancer: Systematic Review of AI-Assisted Endoscopy.” *Diagnostics* 13 (24): 3613. https://doi.org/10.3390/diagnostics13243613.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*.

Lu, Ming Y., Dehan Kong, Jana Lipkova, Richard J. Chen, Rajendra Singh, Drew F. K. Williamson, Tiffany Y. Chen, and Faisal Mahmood. 2020. “Federated Learning for Computational Pathology on Gigapixel Whole Slide Images.” arXiv. https://doi.org/10.48550/arXiv.2009.10190.

Ma, Z., Zhang, M., Liu, J., Yang, A., Li, H., Wang, J., Hua, D., & Li, M. (2022). An Assisted Diagnosis Model for Cancer Patients Based on Federated Learning. Frontiers in Oncology, 12, 860532–860532. https://doi.org/10.3389/fonc.2022.860532

McMahan, H. Brendan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. "Communication-Efficient Learning of Deep Networks from Decentralized Data." arXiv, January 26, 2023. https://doi.org/10.48550/arXiv.1602.05629.

Pati, Sarthak, Ujjwal Baid, Brandon Edwards, Micah Sheller, Shih-Han Wang, G. Anthony Reina, Patrick Foley, et al. 2022. "Federated Learning Enables Big Data for Rare Cancer Boundary Detection." *Nature Communications* 13 (1): 7346. https://doi.org/10.1038/s41467-022-33407-5.

Prayitno, Chi-Ren Shyu, Karisma Trinanda Putra, Hsing-Chung Chen, Yuan-Yu Tsai, K. S. M. Tozammel Hossain, Wei Jiang, and Zon-Yin Shae. 2021. "A Systematic Review of Federated Learning in the Healthcare Area: From the Perspective of Data Properties and Applications." *Applied Sciences* 11 (23): 11191. https://doi.org/10.3390/app112311191.

Qian, Ledan, Libing Hu, Li Zhao, Tao Wang, and Runhua Jiang. "Sequence-Dropout Block for Reducing Overfitting Problem in Image Classification." *IEEE Access* 8 (2020): 62830–40. https://doi.org/10.1109/ACCESS.2020.2983774.

Saeed, Alaa, A. A. Abdel-Aziz, Amr Mossad, Mahmoud A. Abdelhamid, Alfadhl Y. Alkhaled, and Muhammad Mayhoub. "Smart Detection of Tomato Leaf Diseases Using Transfer Learning-Based Convolutional Neural Networks." *Agriculture* 13, no. 1 (2023): 139. https://doi.org/10.3390/agriculture13010139.

Santos, Claudio Filipi Gonçalves Dos, and João Paulo Papa. "Avoiding Overfitting: A Survey on Regularization Methods for Convolutional Neural Networks." *ACM Comput. Surv.* 54, no. 10s (September 13, 2022): 213:1-213:25. https://doi.org/10.1145/3510413.

Sharma, S., Mehra, R. Conventional Machine Learning and Deep Learning Approach for Multi-Classification of Breast Cancer Histopathology Images—a Comparative Insight. J Digit Imaging 33, 632–654 (2020). https://doi.org/10.1007/s10278-019-00307-y

Sheller, Micah J., Brandon Edwards, G. Anthony Reina, Jason Martin, Sarthak Pati, Aikaterini Kotrotsou, Mikhail Milchenko, et al. 2020. "Federated Learning in Medicine: Facilitating Multi-Institutional Collaborations without Sharing Patient Data." *Scientific Reports* 10 (1): 12598. https://doi.org/10.1038/s41598-020-69250-1.

Shorten, Connor, and Taghi M. Khoshgoftaar. 2019. “A Survey on Image Data Augmentation for Deep Learning.” *Journal of Big Data* 6 (1): 60. https://doi.org/10.1186/s40537-019-0197-0.

Song, Zhigang, Shuangmei Zou, Weixun Zhou, Yong Huang, Liwei Shao, Jing Yuan, Xiangnan Gou, et al. 2020. “Clinically Applicable Histopathological Diagnosis System for Gastric Cancer Detection Using Deep Learning.” *Nature Communications* 11 (1): 4294. https://doi.org/10.1038/s41467-020-18147-8.

Sudhakara, M., Y. Vijaya Shambhavi, R. Obulakonda Reddy, N. Badrinath, and K. Reddy Madhavi. “Fish Classification System Using Customized Deep Residual Neural Networks on Small-Scale Underwater Images.” In *Intelligent Computing and Applications*, edited by B. Narendra Kumar Rao, R. Balasubramanian, Shiuh-Jeng Wang, and Richi Nayak, 327–37. Singapore: Springer Nature, 2023. https://doi.org/10.1007/978-981-19-4162-7_31.

Tang, Dehua, Lei Wang, Tingsheng Ling, Ying Lv, Muhan Ni, Qiang Zhan, Yiwei Fu, et al. 2020. “Development and Validation of a Real-Time Artificial Intelligence-Assisted System for Detecting Early Gastric Cancer: A Multicentre Retrospective Diagnostic Study.” *EBioMedicine* 62 (December):103146. https://doi.org/10.1016/j.ebiom.2020.103146.

Teo, Zhen Ling, Liyuan Jin, Nan Liu, Siqi Li, Di Miao, Xiaoman Zhang, Wei Yan Ng, et al. “Federated Machine Learning in Healthcare: A Systematic Review on Clinical Applications and Technical Architecture.” *Cell Reports Medicine* 5, no. 2 (February 2024): 101419. https://doi.org/10.1016/j.xcrm.2024.101419.

Victor Ikechukwu, A., S. Murali, R. Deepu, and R. C. Shivamurthy. “ResNet-50 vs VGG-19 vs Training from Scratch: A Comparative Analysis of the Segmentation and Classification of Pneumonia from Chest X-Ray Images.” *Global Transitions Proceedings*, International Conference on Computing System and its Applications (ICCSA- 2021), 2, no. 2 (November 1, 2021): 375–81. https://doi.org/10.1016/j.gltp.2021.08.027.

Yong, Ming Ping, Yan Chai Hum, Khin Wee Lai, Choon Hian Goh, Wun-She Yap, and Yee Kai Tee. 2023. “Histopathological Gastric Cancer Detection Using Transfer Learning.” In *2023 11th International Conference on Bioinformatics and Computational Biology (ICBCB)*, 123–29. https://doi.org/10.1109/ICBCB57893.2023.10246524.

Yoshida, Hiroshi, Taichi Shimazu, Tomoharu Kiyuna, Atsushi Marugame, Yoshiko Yamashita, Eric Cosatto, Hirokazu Taniguchi, Shigeki Sekine, and Atsushi Ochiai. 2018. “Automated

Histological Classification of Whole-Slide Images of Gastric Biopsy Specimens." *Gastric Cancer* 21 (2): 249–57. https://doi.org/10.1007/s10120-017-0731-8.

Annex A – Model Development

Overfitting in CNN's has always been an issue, particularly as the depth of network increases when being applied to smaller Data samples. The small data sample (single clinic) combined with low contrast imagery make it increasingly difficult to develop a durable model. Overfitting occurs when the model performs well on the training data then poorly on the testing data; It is usually caused by network complexity or a lack of training data (Quian et al, 2020). While overfitting can be a significant problem there are several regularisation techniques that have been developed to combat this issue. Santos et al (2022) published a paper on several techniques that have been found to improve the performance of CNNs through regularisation of the network. Santos et al (2022) segmented the techniques into three broad groups:

- Alterations to input – data augmentation
- Alterations to the network architecture
- Alterations to the label

Examples of all three of these regularisation techniques were applied to the development of the CNNs either as part of transfer learning or as part of model development.

Data Augmentation

PyTorch offers several methods that can be applied to augment data as part of developing the Data set for use in the model training. The following table highlights the differences based on a single model with augmentation techniques applied and the accuracy of the model under controlled conditions.

Experiment Ref	**Augmentation Technique**	**Model Accuracy**
Base	No Augmentation	F1 = 0.96 Loss = 0.103
Aug 1	Colour Jitter (brightness 0.5, contrast 0.3)	F1 = 0.928 Loss = .215
Aug 2	Colour Jitter (contrast 0.8)	F1 = 0.958 Loss = 0.149

These experiments were conducted over 30 epochs, as such there is possibility that the model could continue to improve over more epochs. These experiments were conducted on the model with dropout in the convolutions – structure outlined below.

Alterations to Network Architecture

Rather than use more of the data or conduct complex and detailed preprocessing of the data it was decided that an attempt should be made to develop a custom model that had a greater amount of regularisation applied in the convolutional networks themselves. The resulting model was based on a RESNET architecture. RESNET architecture was used to keep the model *lightweight* relative to the other model architectures as the ability to fully train a model in a federated setting would take a significant amount of time and compute. Park & Nujon (2017) served as the motivation for the experiment. The resulting model structure is below:

Layer (type)	Output Shape	Param #
Conv2d-1	[-1, 64, 80, 80]	1,728
BatchNorm2d-2	[-1, 64, 80, 80]	128
ReLU-3	[-1, 64, 80, 80]	0
MaxPool2d-4	[-1, 64, 40, 40]	0
Conv2d-5	[-1, 64, 40, 40]	36,864
BatchNorm2d-6	[-1, 64, 40, 40]	128
ReLU-7	[-1, 64, 40, 40]	0
Dropout-8	[-1, 64, 40, 40]	0
Conv2d-9	[-1, 64, 40, 40]	36,864
BatchNorm2d-10	[-1, 64, 40, 40]	128
ReLU-11	[-1, 64, 40, 40]	0
Dropout-12	[-1, 64, 40, 40]	0
BasicBlock-13	[-1, 64, 40, 40]	0
Conv2d-14	[-1, 64, 40, 40]	36,864
BatchNorm2d-15	[-1, 64, 40, 40]	128
ReLU-16	[-1, 64, 40, 40]	0
Dropout-17	[-1, 64, 40, 40]	0
Conv2d-18	[-1, 64, 40, 40]	36,864
BatchNorm2d-19	[-1, 64, 40, 40]	128
ReLU-20	[-1, 64, 40, 40]	0
Dropout-21	[-1, 64, 40, 40]	0
BasicBlock-22	[-1, 64, 40, 40]	0
Conv2d-23	[-1, 128, 20, 20]	73,728
BatchNorm2d-24	[-1, 128, 20, 20]	256
ReLU-25	[-1, 128, 20, 20]	0
Dropout-26	[-1, 128, 20, 20]	0

Conv2d-27	[-1, 128, 20, 20]	147,456
BatchNorm2d-28	[-1, 128, 20, 20]	256
Conv2d-29	[-1, 128, 20, 20]	8,192
BatchNorm2d-30	[-1, 128, 20, 20]	256
ReLU-31	[-1, 128, 20, 20]	0
Dropout-32	[-1, 128, 20, 20]	0
BasicBlock-33	[-1, 128, 20, 20]	0
Conv2d-34	[-1, 128, 20, 20]	147,456
BatchNorm2d-35	[-1, 128, 20, 20]	256
ReLU-36	[-1, 128, 20, 20]	0
Dropout-37	[-1, 128, 20, 20]	0
Conv2d-38	[-1, 128, 20, 20]	147,456
BatchNorm2d-39	[-1, 128, 20, 20]	256
ReLU-40	[-1, 128, 20, 20]	0
Dropout-41	[-1, 128, 20, 20]	0
BasicBlock-42	[-1, 128, 20, 20]	0
Conv2d-43	[-1, 256, 10, 10]	294,912
BatchNorm2d-44	[-1, 256, 10, 10]	512
ReLU-45	[-1, 256, 10, 10]	0
Dropout-46	[-1, 256, 10, 10]	0
Conv2d-47	[-1, 256, 10, 10]	589,824
BatchNorm2d-48	[-1, 256, 10, 10]	512
Conv2d-49	[-1, 256, 10, 10]	32,768
BatchNorm2d-50	[-1, 256, 10, 10]	512
ReLU-51	[-1, 256, 10, 10]	0
Dropout-52	[-1, 256, 10, 10]	0
BasicBlock-53	[-1, 256, 10, 10]	0
Conv2d-54	[-1, 256, 10, 10]	589,824
BatchNorm2d-55	[-1, 256, 10, 10]	512
ReLU-56	[-1, 256, 10, 10]	0
Dropout-57	[-1, 256, 10, 10]	0
Conv2d-58	[-1, 256, 10, 10]	589,824
BatchNorm2d-59	[-1, 256, 10, 10]	512
ReLU-60	[-1, 256, 10, 10]	0
Dropout-61	[-1, 256, 10, 10]	0
BasicBlock-62	[-1, 256, 10, 10]	0
Conv2d-63	[-1, 512, 5, 5]	1,179,648
BatchNorm2d-64	[-1, 512, 5, 5]	1,024
ReLU-65	[-1, 512, 5, 5]	0

Dropout-66	[-1, 512, 5, 5]	0
Conv2d-67	[-1, 512, 5, 5]	2,359,296
BatchNorm2d-68	[-1, 512, 5, 5]	1,024
Conv2d-69	[-1, 512, 5, 5]	131,072
BatchNorm2d-70	[-1, 512, 5, 5]	1,024
ReLU-71	[-1, 512, 5, 5]	0
Dropout-72	[-1, 512, 5, 5]	0
BasicBlock-73	[-1, 512, 5, 5]	0
Conv2d-74	[-1, 512, 5, 5]	2,359,296
BatchNorm2d-75	[-1, 512, 5, 5]	1,024
ReLU-76	[-1, 512, 5, 5]	0
Dropout-77	[-1, 512, 5, 5]	0
Conv2d-78	[-1, 512, 5, 5]	2,359,296
BatchNorm2d-79	[-1, 512, 5, 5]	1,024
ReLU-80	[-1, 512, 5, 5]	0
Dropout-81	[-1, 512, 5, 5]	0
BasicBlock-82	[-1, 512, 5, 5]	0
AdaptiveAvgPool2d-83	[-1, 512, 1, 1]	0
Linear-84	[-1, 2]	1,026
ResNet18-85	[-1, 2]	0

==========

Total params: 11,169,858
Trainable params: 11,169,858
Non-trainable params: 0

Input size (MB): 0.07
Forward/backward pass size (MB): 37.89
Params size (MB): 42.61
Estimated Total Size (MB): 80.58

By applying the dropout after the activation function, it effectively hinders the model's learning in a small way inside the neural network itself. This is combated somewhat with the inherent design of the skip layers in RESNET architecture. The outcome likely means that it will take longer to train the model, but the model will be more durable in a federated setting as it will have access to more data across the clinics. Again, the focus was making it deployable in a real-world scenario.

Alterations to the Labels

MixUp was one technique used to alter the labels. This was first implemented by Zhang et al (2017). The procedure trains neural networks on convex combinations of pairs of the labels and images. In doing this Mixup regularises the network making it more robust as it reduces memorisation thereby reducing the likelihood of the model overfitting. Mixup was applied to the model outlined above.

CutMix was the other method used to alter the labels and the image itself. CutMix was first implemented in the paper by Yun et al (2019). The technique involves the cutting and mixing of patches and labels across batches of images as part of the training process. The original paper highlights that CutMix consistently outperforms other regularisation strategies; furthermore, in this context it makes the model more robust to input corruptions.

The results for the two methods are outlined below:

Experiment Ref	**Augmentation Technique**	**Model Accuracy**
Base	No Augmentation	F1 = 0.965 Loss = 0.103
Aug 3	MixUp	F1 = 0.962 Loss = 0.189
Aug 4	CutMix	F1 = 0.972 Loss = 0.118

Conclusion

When compared with other regularisation techniques and, to keep the initial model and the process around it as simple as possible the decision to use the model architecture as the primary source of regularisation to combat overfitting was made. In the future as the FL process is better understood following deployment then these other techniques could be applied to the process to further enhance the quality of the model.

ANEX B – Base Model Selection

Transfer Learning Process

The transfer learning was an iterative process as the very act of fine tuning a CNN on a specific data set can be complex. The three models trained initially were RESNET18, VGG16 and VGG19. During experimentation it was discovered that the models were overfitting in i.e they would perform well in training but the results on validation and test data were poor. This led the team to develop a custom model. The details of the custom model are at annex A.

Alter the Classifier in each of the models

The process of transfer learning involves altering the architecture of the model slightly to make it suitable for use in our specific problem. The feature extraction layers for each of the models remained the same and only the fully connected classifiers at the end of each of the models was adjusted. All of the pretrained models were trained on the ImageNet data set with 1000 classes of images the classifier had to be altered for a binary task (two classes). Each of the models was adjusted a minimally as possible to preserve the original intent behind the model development.

Overfitting

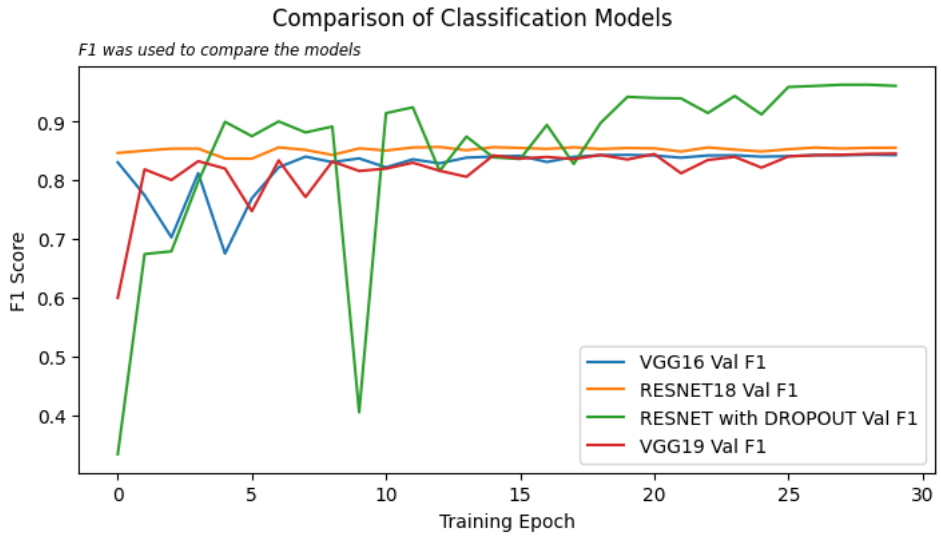
The RESNET architecture was developed in response to the vanishing and exploding curse that hunts most CNN architectures. The skip connections allow the model to feed forward features learned as part of training – effectively skipping some connections. What this does is avoid the passing of ever increasing poor gradients to the future connections. Effectively they are normalised by the feature extracted a few connections earlier and it bring a sense of 'normal' to the learning process.

All models used for transfer learning had a level of dropout in the final layers of the model – the fully connected classification module. This has been found to be effective in assisting to counter overfitting as well as poor image quality in the training data (Saeed et al, 2023). In addition to controlling for poor image quality Klanecek et al (2023) found that by inserting dropout layers into the model architecture both during training and testing that the results were considerably better than without dropout in the model architecture (>2.5% better). A similar result was attained by Sudhakara et al (2023) while keeping the size of the model at an acceptable number of trainable parameters - a key consideration when simulating FL environments.

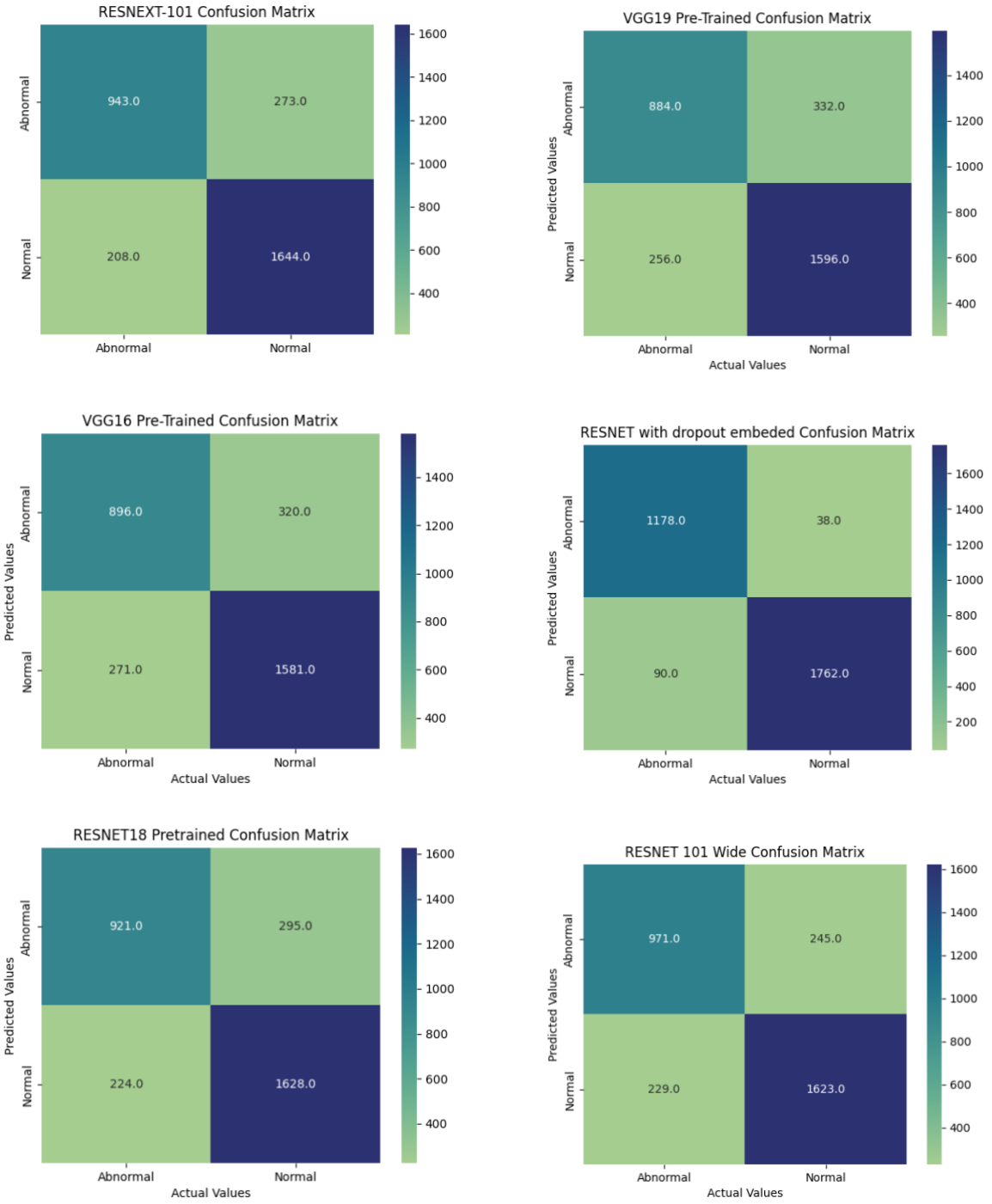
Transfer Learning Results

Results on the Data were mixed across the models – not state-of-the-art when compared to other efforts on the GasHisSDB. There was a trade-off between developing a first-class model and something that was usable in a real-world context. The amount of *compute* required to apply a deep and complex model in a FL context is significant when being run in simulated context. Simple yet effective model architectures were preferred. As part of the experimentation the group found that using pre-trained models the results didn't perform as expected. One element that emerged as the number of training epochs increased was the issue of over-fitting – this led to the development of the custom model. The details of the model development and other options to prevent overfitting are at Annex A.

The initial transfer learning training was done over 30 epochs on a single clinic of data. The comparison of the different models during the validation on a single clinic of data is below. While the RESNET with drop-out appears less stable during the early stages of training it ultimately surpasses the other pre-trained models.



The confusion matrices below highlight the differences in the performance of the models. The values in the confusion matrices were taken from the test data from a single clinic.



The results from the pre-trained models are comparable. The results from the custom model are an order of magnitude better than the pre-trained models. These results were thought to be due to the model being fully trained. As such a test was done and both the VGG16 and the RESNET18 models were fully trained on the same set of data. The results didn't vary greatly from those above and the RESNET architecture with dropout still outperformed all other models for this task.

For instance, the dataset consists of 97,076 abnormal and 148,120 normal images of patients respectively. GasHisSDB contains pictures in PNG format obtained by electron microscopy.

Normal images contain no cancerous regions, with cells showing little to no atypia, regular single-layer arrangement, and minimal mitosis. Therefore, if no signs of cellular or tissue abnormality, such as structural alterations, discoloration, or atypical formations, are observed and all defining characteristics of a normal pathological image are present, the image can be classified as normal. These characteristics make them easily identifiable as normal under a microscope, allowing whole images to be directly cropped for dataset creation. (https://www.sciencedirect.com/science/article/pii/S0010482521010015?via%3Dihub)

As gastric cancer progresses, cancerous nests expand, infiltrating sequentially from the mucosal layer to the muscle layer and eventually reaching the serosal layer, giving the tissue a hard texture and a characteristic gray-white appearance on section. Microscopically, cancer cells may display diverse formations, including nest, acinar, tubular, or cord-like structures, often with a distinct boundary between the tumor cells and the surrounding stroma. However, once the cancer cells infiltrate into the stroma, this boundary becomes indistinct. When cells exhibit glandular or adenoid structures of varying sizes, irregular shapes, and atypical arrangements, the image can be classified as abnormal. (https://www.sciencedirect.com/science/article/pii/S0010482521010015?via%3Dihub)

In such abnormal images, cancer cells are frequently arranged in disorganized, multilayered formations with nuclei of varying sizes, accompanied by evidence of cellular division abnormalities. During the creation of the abnormal image dataset, each cancerous region is cropped based on the ground truth (GT) from the original images. Images are further filtered to ensure a high concentration of cancerous areas, typically at least 50%, to emphasize significant pathological regions. (https://www.sciencedirect.com/science/article/pii/S0010482521010015?via%3Dihub)

Histological analysis through microscopy remains the gold standard for diagnosing and staging cancer, providing pathologists with critical insights into cellular morphology and spatial organisation. In this process, tissue samples are prepared on slides or whole slide images, which pathologists meticulously examine to identify distinctive cellular features, such as shape, size, arrangement, and spatial relationships within the tissue structure. These microscopic details are essential for determining the cancer's type, stage, and aggressiveness, guiding decisions on the appropriate course of treatment. The analysis captures subtle changes in cell morphology that often indicate the extent of tumor differentiation and infiltration, providing valuable information on prognosis. As the cornerstone of cancer diagnostics, histological examination enables a detailed assessment

of cancerous tissues, hence stained nuclei contain purplish-blue and cytoplasm pink, making cellular structures easily distinguishable. Meanwhile, normal images display more pink and white areas, while abnormal ones have more disorganised purplish-blue regions (Hu et al, 2022).

(https://pmc.ncbi.nlm.nih.gov/articles/PMC6095898/#:~:text=Thus%2C%20the%20diagnostic%20strategy%20based,an%20accurate%20clinical%20tumor%20diagnosis and https://www.sciencedirect.com/science/article/pii/S0344033823003941)

References

Park, Sungheon, and Nojun Kwak. 2017. "Analysis on the Dropout Effect in Convolutional Neural Networks." In *Computer Vision – ACCV 2016*, edited by Shang-Hong Lai, Vincent Lepetit, Ko Nishino, and Yoichi Sato, 189–204. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-54184-6_12.

Santos, Claudio Filipi Gonçalves Dos, and João Paulo Papa. 2022. "Avoiding Overfitting: A Survey on Regularization Methods for Convolutional Neural Networks." *ACM Comput. Surv.* 54 (10s): 213:1-213:25. https://doi.org/10.1145/3510413.

Yun, Sangdoo, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. 2019. "CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features." arXiv. http://arxiv.org/abs/1905.04899.

Zhang, Hongyi, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. "Mixup: Beyond Empirical Risk Minimization." arXiv. http://arxiv.org/abs/1710.09412.