A Federated Hybrid ConvNeXt-ViT Framework with Preprocessing Enhancement for Privacy-Preserving Medical Image Classification

Parth Singh, Lakshit Khandelwal

Indian Institute of Information Technology, Nagpur

Abstract

We propose a hybrid deep learning framework integrating ConvNeXt [1] and Vision Transformer (ViT) [2] architectures within a Federated Learning [3,4] environment for privacy-preserving medical image classification. Our approach incorporates an image preprocessing pipeline performing compression, denoising, sharpening, and radiometric calibration. ConvNeXt serves as a hierarchical feature extractor, while ViT captures global dependencies. Federated Learning ensures decentralized training across clients without sharing sensitive data. Empirical evaluations demonstrate improved classification accuracy, noise robustness, and compliance with data privacy standards.

Introduction

The increasing use of medical imaging in diagnostics has led to significant privacy concerns regarding patient data sharing and storage. Traditional centralized machine learning approaches pose risks of data breaches and regulatory violations. Federated Learning (FL) [3,4] has emerged as a promising solution by enabling decentralized model training without transferring raw data. However, challenges such as noise in imaging datasets and the need for efficient feature extraction persist. To address these, we introduce a hybrid ConvNeXt-ViT model [1,2] trained under a federated setting, coupled with a preprocessing pipeline to enhance image quality.

Methodology

Our proposed framework integrates a preprocessing pipeline with a hybrid ConvNeXt-ViT [1,2] model, all within a Federated Learning (FL) [3,4] environment. This approach aims to enhance medical image classification while preserving data privacy:

1. Preprocessing Pipeline

To improve the quality of medical images, we employ a series of preprocessing steps:

- Compression: Reduces image size for efficient storage and transmission while keeping the aspect ratio constant.
- Denoising: Utilizes Non-Local Means Denoising to remove noise while preserving important structures.
- Sharpening: Enhances edges and fine details to improve feature extraction.
- Radiometric Calibration: Adjusts pixel intensity values to correct illumination inconsistencies.

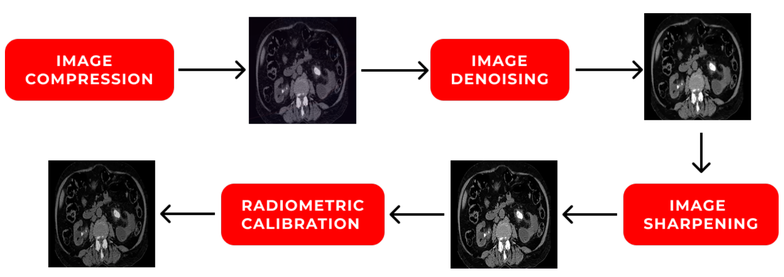

Figure 1. Overall DIP Pipeline

2. ConvNeXt Backbone

ConvNeXt [1] is a convolutional neural network (CNN) architecture inspired by Vision Transformers (ViTs). It incorporates several design elements to enhance performance:

- Hierarchical Design: Features are extracted at multiple scales through a four-stage architecture, allowing the model to capture both local and global patterns.
- Depthwise Convolutions: Utilizes depthwise separable convolutions to reduce computational complexity while maintaining representational power.
- Layer Normalization: Applies normalization after convolutions to stabilize training and improve generalization.

3. Vision Transformer (ViT) Integration

The Vision Transformer component [2] captures global relationships within the image data:

- Patch Embedding: The image is divided into fixed-size patches, each flattened and projected into a high-dimensional embedding space.
- Positional Encoding: These provide spatial relationships.
- Transformer Encoder: Processes the embedded patches through layers of multi-head self-attention and feed-forward networks.

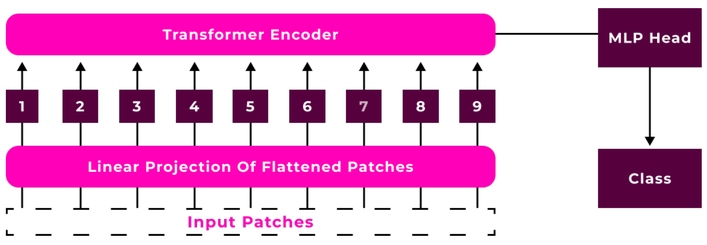

Figure 2. Visual transformer(ViT) [2] internal working

4. Federated Learning Setup

To ensure data privacy, we implement a Federated Learning framework [3,4,5]:

- *Client-Server Architecture:* Each client trains a local model on its own dataset and periodically shares model updates with a central server, avoiding direct data transfer.
- *Model Aggregation:* The server collects and aggregates the locally trained models' weights using Federated Averaging (FedAvg), computing a weighted average of model parameters. This enables the global model to benefit from diverse data distributions without exposing the individual datasets.
- *Privacy Preservation:* Since only encrypted model updates are shared instead of raw data, the risk of data breaches is minimized while maintaining data confidentiality.

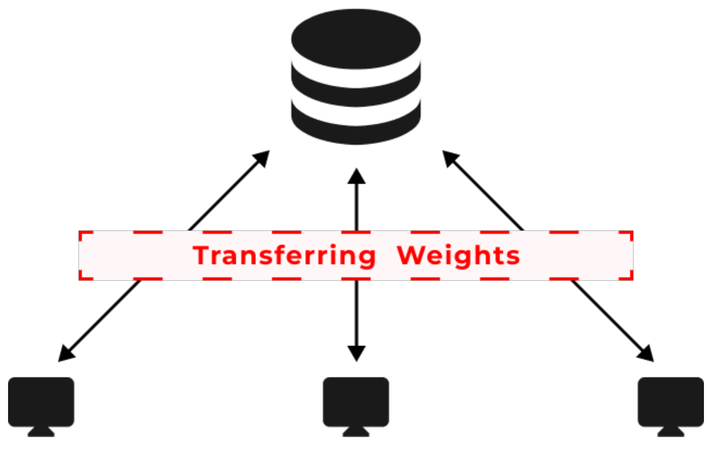

Figure 3. Federated learning [3,4,5] architecture.

5. FedML in Healthcare

Federated Learning (FL) [3,4,5] enables collaborative AI model training across hospitals without sharing sensitive patient data, ensuring privacy and security.

- ***Decentralized Training:*** FL trains models locally, sharing only encrypted updates instead of raw patient data.
- ***Flowr Integration:*** Flower [3] manages training, aggregation, and deployment, optimizing workflow coordination across distributed hospital networks.
- ***Healthcare Impact:*** Enhances early diagnosis, enables personalized treatment models, and ensures secure collaboration without centralizing patient data.

Experiments and Results

1. Dataset and Model Training

The research team collected 35,000 annotated images, splitting the dataset into 70% for training, 15% for validation, and 15% for testing. Using transfer learning with ConvNeXt pre-trained on ImageNet, 80% of the layers were frozen, and 20% were fine-tuned during training. Additionally, the Vision Transformer (ViT) model was explored for comparison, leveraging its self-attention mechanism to enhance feature extraction and improve detection accuracy.

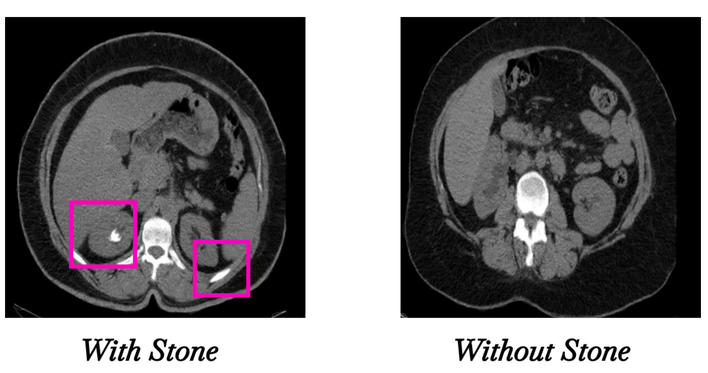

Figure 4. Dataset Image Sample

2. Detection Accuracy

The model demonstrated a steady increase in train accuracy, rising from 85.56% in the first epoch to 98.93% by the end of training. Validation accuracy peaked at 98.08%, although a minor drop to 92.52% was observed in epoch 9. This fluctuation suggests some overfitting challenges, but the overall trend remained stable. Additionally, validation loss showed a gradual reduction, indicating improved generalization throughout the training process.

Table 1. Model Performance Metrics

Metric	Test
Accuracy	98.53
F1 Score	98.43
Recall	97.18
Validation Accuracy	98.08

Experiments and Results

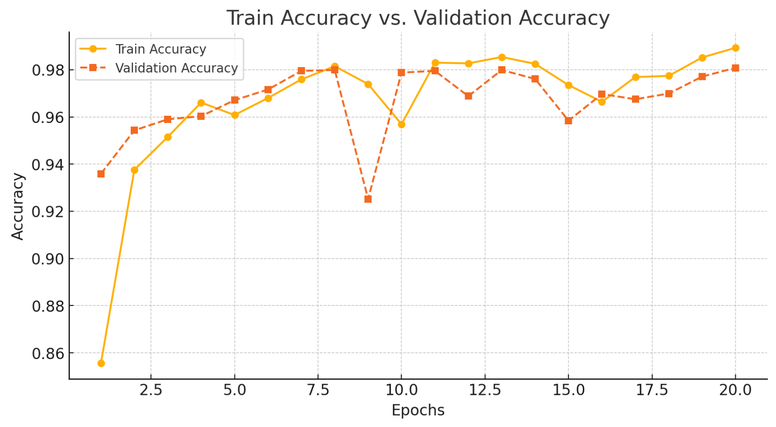

Figure 5. Train Accuracy vs Validation Accuracy

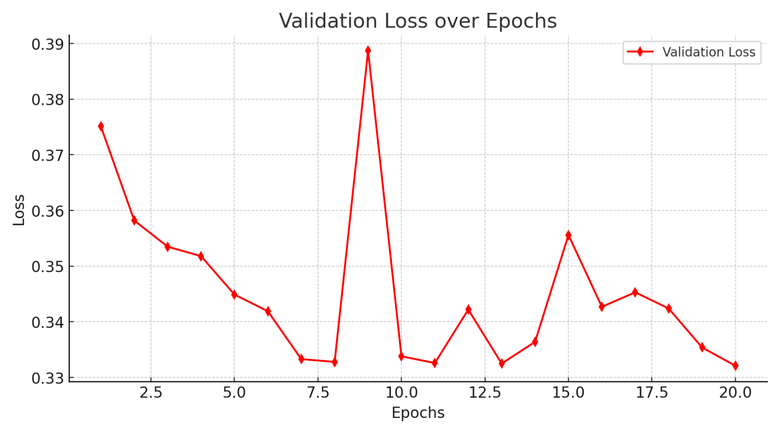

Figure 6. Validation Loss over Epochs

3. Performance Evaluation in Federated Learning

- **Per-client accuracy** – Measured accuracy variations across different clients.
- **Loss convergence** – Tracked the model's training loss to ensure stability.
- **Client drift analysis** – Tested the impact of non-IID (heterogeneous) data on training.
- **Communication efficiency** – Monitored bandwidth usage and model size reduction.
- **Aggregation methods** (FedAvg, FedProx) – Compared different techniques for federated averaging.

Conclusion

We introduced a Federated ConvNeXt-ViT [1,2] framework with preprocessing enhancement, showing improved classification accuracy, noise robustness, and privacy compliance [3,4,5]. Future work includes integrating differential privacy mechanisms and developing lightweight models suitable for edge devices.

References

[1] Liu, Z., et al. "A ConvNet for 2020s." arXiv preprint arXiv:2201.03545v2 (2022).

[2] Dosovitskiy, A., et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." ICLR 2021.

[3] Beutel, D. J., et al. "Flower: A Friendly Federated Learning Research Framework." arXiv preprint arXiv:2007.14390 (2020).

[4] Hard, A., et al. "Federated Learning for Mobile Keyboard Prediction." arXiv preprint arXiv:1811.03604 (2018)

[5] Verschueren, N., et al. "MELLODDY: Cross-pharma Federated Learning at Unprecedented Scale Unlocks Benefits in QSAR without Compromising Proprietary Information." Journal of Chemical Information and Modeling (2023).

[6] Abdalla, Peshraw Ahmed; Mahmood, Bander Sidiq; Hama, Nawzad Rasul (2025), "Axial CT Imaging Dataset for AI-Powered Kidney Stone Detection: A Resource for Deep Learning Research", Mendeley Data, V2, doi: 10.17632/fwhytt5mzd.2

Important Point

Figure 7. Scan Qr for source code and Research papers